White Paper

Integrate Cisco UCS C240 M4 Rack Server with NVIDIA GRID Graphics Processing Unit Card on VMware Horizon View 6.1

Contents

What You Will Learn

With the increased processor power of today's Cisco UCS® B-Series Blade Servers and C-Series Rack Servers, applications with demanding graphics components are now being considered for virtualization. To enhance the capability to deliver these high-performance and graphics-intensive applications, Cisco includes the NVIDIA GRID K1 and K2 cards in the Cisco Unified Computing System™ (Cisco UCS) portfolio of PCI Express (PCIe) cards for the C-Series Rack Servers.

With the addition of the new graphics processing capabilities, the engineering, design, imaging, and marketing departments of organizations can now experience the benefits that desktop virtualization brings to these applications.

This new graphics capability enables organizations to centralize their graphics workloads and data in the data center. This capability greatly benefits organizations that need to be able to shift work geographically. Until now, graphics files have been too large to move, and the files have had to be local to the person using them to be usable.

The PCIe graphics cards in the Cisco UCS C-Series offer these benefits:

- Support for full-length, full-power NVIDIA GRID cards in a 2-rack-unit (2RU) form factor
- Cisco UCS Manager integration for management of the servers and GRID cards
- End-to-end integration with the Cisco UCS management solutions, including Cisco UCS Central Software and Cisco UCS Director
- Cisco UCS C240 M4 Rack Servers with two NVIDIA GRID cards provide more efficient rack space than the 2-slot, 2.5-inch equivalent rack unit: the HP WS460c workstation blade with the NVIDIA GRID card in a second slot

The purpose of this document is to help our partners and customers integrate NVIDIA GRID graphics processing cards and the C240 M4 servers on VMware vSphere and VMware Horizon in virtual dedicated graphics acceleration (vDGA), virtual shared graphics acceleration (vSGA), and virtual graphics processing unit (vGPU) modes.

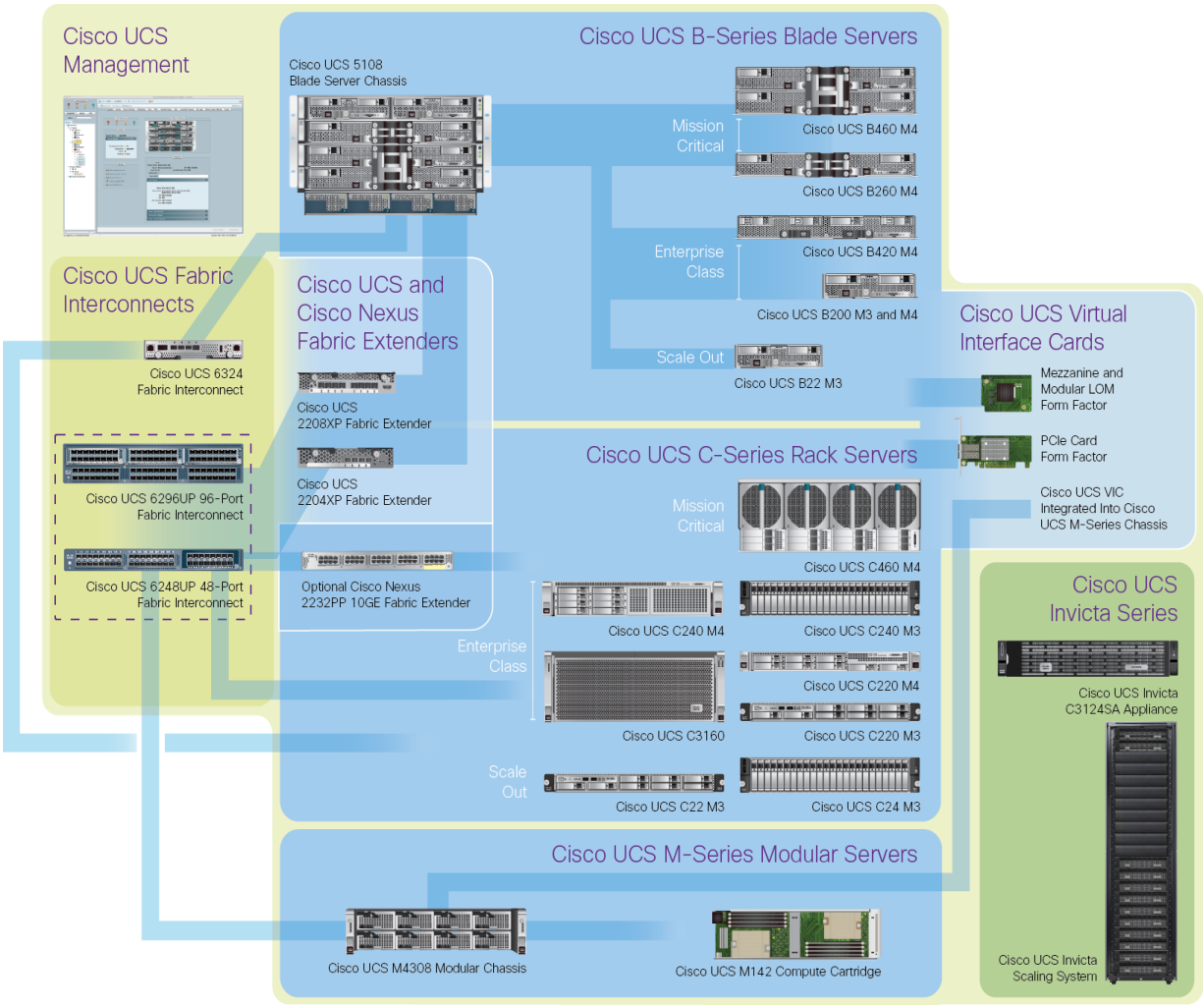
Please contact our partners, NVIDIA and VMware, for lists of applications that are supported by the card, hypervisor, and desktop broker in each mode.

Our objective is to provide the reader with specific methods for integrating C240 M4 servers with GRID K1 and K2 cards with VMware products so that the servers, hypervisor, and desktop broker are ready for installation of graphics applications.

Cisco Unified Computing System

Cisco UCS is a next-generation data center platform that unites computing, networking, and storage access. The platform, optimized for virtual environments, is designed using open industry-standard technologies and aims to reduce total cost of ownership (TCO) and increase business agility. The system integrates a low-latency; lossless 10 Gigabit Ethernet unified network fabric with enterprise-class, x86-architecture servers. It is an integrated, scalable, multichassis platform in which all resources participate in a unified management domain.

Figure 1. Cisco UCS Components



The main components of Cisco UCS (Figure 1) are:

- **Computing:** The system is based on an entirely new class of computing system that incorporates blade servers based on Intel® Xeon® processor E5-2600/4600 v3 and E7-2800 v3 family CPUs.
- **Network:** The system is integrated onto a low-latency, lossless, 10-Gbps unified network fabric. This network foundation consolidates LANs, SANs, and high-performance computing (HPC) networks, which are separate networks today. The unified fabric lowers costs by reducing the number of network adapters, switches, and cables, and by decreasing the power and cooling requirements.
- **Virtualization:** The system unleashes the full potential of virtualization by enhancing the scalability, performance, and operational control of virtual environments. Cisco security, policy enforcement, and diagnostic features are now extended into virtualized environments to better support changing business and IT requirements.
- **Storage access:** The system provides consolidated access to local storage, SAN storage, and network-attached storage (NAS) over the unified fabric. With storage access unified, Cisco UCS can access storage

over Ethernet, Fibre Channel, Fibre Channel over Ethernet (FCoE), and Small Computer System Interface over IP (iSCSI). This capability provides customers with choice for storage access and investment protection. In addition, server administrators can preassign storage-access policies for system connectivity to storage resources, simplifying storage connectivity and management and helping increase productivity.

- **Management:** Cisco UCS uniquely integrates all system components, enabling the entire solution to be managed as a single entity by Cisco UCS Manager. The manager has an intuitive GUI, a command-line interface (CLI), and a robust API for managing all system configuration processes and operations.

Cisco UCS is designed to deliver:

- Reduced TCO and increased business agility
- Increased IT staff productivity through just-in-time provisioning and mobility support
- Cohesive, integrated system that unifies the technology in the data center; the system is managed, serviced, and tested as a whole
- Scalability through a design for hundreds of discrete servers and thousands of virtual machines and the capability to scale I/O bandwidth to match demand
- Industry standards supported by a partner ecosystem of industry leaders

Cisco UCS Manager

Cisco UCS Manager provides unified, embedded management of all software and hardware components of Cisco UCS through an intuitive GUI, a CLI, and an XML API. Cisco UCS Manager provides a unified management domain with centralized management capabilities and can control multiple chassis and thousands of virtual machines.

Cisco UCS Fabric Interconnect

The Cisco UCS 6200 Series Fabric Interconnects are a core part of Cisco UCS, providing both network connectivity and management capabilities for the system. The 6200 Series offers line-rate, low-latency, lossless 10 Gigabit Ethernet, FCoE, and Fibre Channel functions.

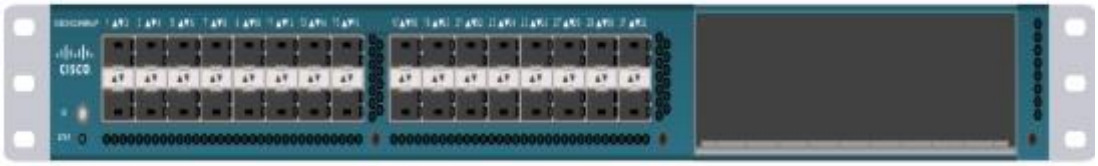
The Cisco UCS 6200 Series provides the management and communication backbone for the Cisco UCS B-Series Blade Servers and Cisco UCS 5100 Series Blade Server Chassis. All chassis, and therefore all blades, attached to the fabric interconnects become part of a single, highly available management domain. In addition, by supporting unified fabric, the 6200 Series provides both the LAN and SAN connectivity for all blades within the domain.

For networking, the 6200 Series uses a cut-through architecture, supporting deterministic, low-latency, line-rate 10 Gigabit Ethernet on all ports, 1-terabit (Tb) switching capacity, and 160 Gbps of bandwidth per chassis, independent of packet size and enabled services. The product family supports Cisco low-latency, lossless, 10 Gigabit Ethernet unified network fabric capabilities, increasing the reliability, efficiency, and scalability of Ethernet networks. The fabric interconnects support multiple traffic classes over a lossless Ethernet fabric from the blade server through the interconnect. Significant TCO savings come from an FCoE-optimized server design in which network interface cards (NICs), host bus adapters (HBAs), cables, and switches can be consolidated.

Cisco UCS 6248UP Fabric Interconnect

The Cisco UCS 6248UP 48-Port Fabric Interconnect (Figure 2) is a 1RU 10 Gigabit Ethernet, FCoE, and Fibre Channel switch offering up to 960-Gbps throughput and up to 48 ports. The switch has thirty-two 1- and 10-Gbps fixed Ethernet, FCoE, and Fibre Channel ports and one expansion slot.

Figure 2. Cisco UCS 6248UP Fabric Interconnect



Cisco UCS C-Series Rack Servers

Cisco UCS C-Series Rack Servers keep pace with Intel Xeon processor innovation by offering the latest processors with an increase in processor frequency and improved security and availability features. With the increased performance provided by the Intel Xeon processor E5-2600 and E5-2600 v2 product families, C-Series servers offer an improved price-to-performance ratio. They also extend Cisco UCS innovations to an industry-standard rack-mount form factor, including a standards-based unified network fabric, Cisco VN-Link virtualization support, and Cisco Extended Memory Technology.

Designed to operate both in standalone environments and as part of Cisco UCS, these servers enable organizations to deploy systems incrementally—using as many or as few servers as needed—on a schedule that best meets the organization's timing and budget. C-Series servers offer investment protection through the capability to deploy them either as standalone servers or as part of Cisco UCS.

One compelling reason that many organizations prefer rack-mount servers is the wide range of I/O options available in the form of PCIe adapters. C-Series servers support a broad range of I/O options, including interfaces supported by Cisco as well as adapters from third parties.

Cisco UCS C240 M4 Rack Server

The Cisco UCS C240 M4 (Figures 3 and 4 and Table 1) is designed for both performance and expandability over a wide range of storage-intensive infrastructure workloads, from big data to collaboration. The enterprise-class C240 M4 server further extends the capabilities of the Cisco UCS portfolio in a 2RU form factor with the addition of the Intel Xeon processor E5-2600 and E5-2600 v2 product families, which deliver an outstanding combination of performance, flexibility, and efficiency gains.

The C240 M4 offers up to two Intel Xeon processor E5-2600 or E5-2600 v2 CPUs, 24 DIMM slots, 24 disk drives, and four 1 Gigabit Ethernet LAN-on-motherboard (LOM) ports to provide exceptional levels of internal memory and storage expandability and exceptional performance.

The C240 M4 interfaces with the Cisco UCS virtual interface card (VIC). The VIC is a virtualization-optimized FCoE PCIe 2.0 x8 10-Gbps adapter designed for use with C-Series Rack Servers. The VIC is a dual-port 10 Gigabit Ethernet PCIe adapter that can support up to 256 PCIe standards-compliant virtual interfaces, which can be dynamically configured so that both their interface type (NIC or HBA) and identity (MAC address and worldwide name [WWN]) are established using just-in-time provisioning. An additional five PCIe slots are available for

certified third-party PCIe cards. The server is equipped to handle 24 on-board serial-attached SCSI (SAS) or solid-state disk (SSD) drives along with shared storage solutions offered by our partners.

The C240 M4 server's disk configuration delivers balanced performance and expandability to best meet individual workload requirements. With up to 12 large form factor (LFF) or 24 small form factor (SFF) internal drives, the C240 M4 optionally offers 10,000- and 15,000-rpm SAS drives to deliver a high number of I/O operations per second (IOPS) for transactional workloads such as database management systems. In addition, high-capacity SATA drives provide an economical, large-capacity solution. Superfast SSD drives are a third option for workloads that demand extremely fast access to smaller amounts of data. A choice of RAID controller options also helps increase disk performance and reliability.

The C240 M4 further increases performance and customer choice over many types of storage-intensive applications such as:

- Collaboration
- Small and medium-sized business (SMB) databases
- Big data infrastructure
- Virtualization and consolidation
- Storage servers
- High-performance appliances

The C240 M4 can be deployed as a standalone server or as part of Cisco UCS, which unifies computing, networking, management, virtualization, and storage access into a single integrated architecture that enables end-to-end server visibility, management, and control in both bare-metal and virtualized environments. Within a Cisco UCS deployment, the C240 M4 takes advantage of Cisco's standards-based unified computing innovations, which significantly reduce customers' TCO and increase business agility.

For more information about the Cisco UCS C240 M4 Rack Server, see:

- http://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-c240-m4-rack-server/index.html
- http://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c240-m4-rack-server/datasheet-c78-732455.html

Figure 3. Cisco UCS C240 M4 Rack Server Front View

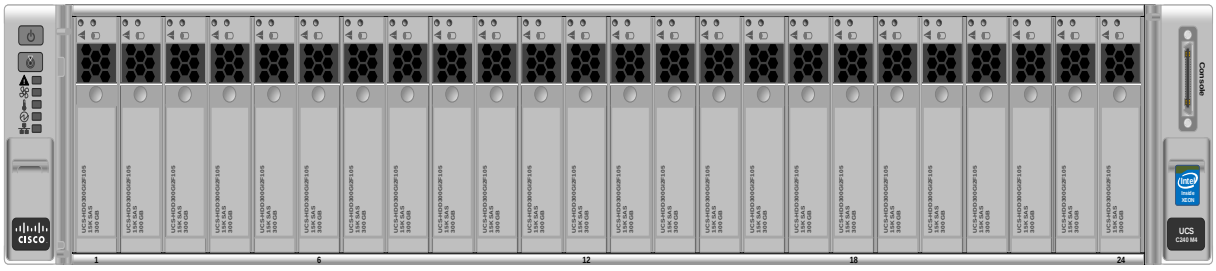


Figure 4. Cisco UCS C240 M4 Rack Server Rear View

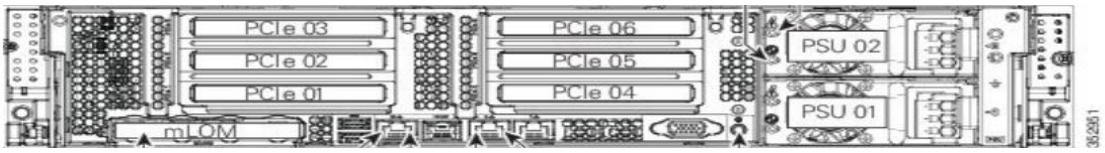


Table 1. Cisco UCS C240 M4 PCIe Slots

PCIe Slot	Length	Lane
1	3/4	x8
2	**Full**	**X16**
3	Full	X8
4	3/4	X8
5	**Full**	**X16**
6	Full	X8

Cisco UCS VIC 1227 Modular LOM

A Cisco innovation, the Cisco UCS VIC 1227 (Figures 5 and 6) is a dual-port, Enhanced Small Form-Factor Pluggable (SFP+), 10 Gigabit Ethernet and FCoE–capable, PCIe modular LAN on motherboard (mLOM) adapter. It is designed exclusively for the M4 generation of Cisco UCS C-Series Rack Servers and for the Cisco UCS C3160 Rack Server, which provides dense storage capacity.

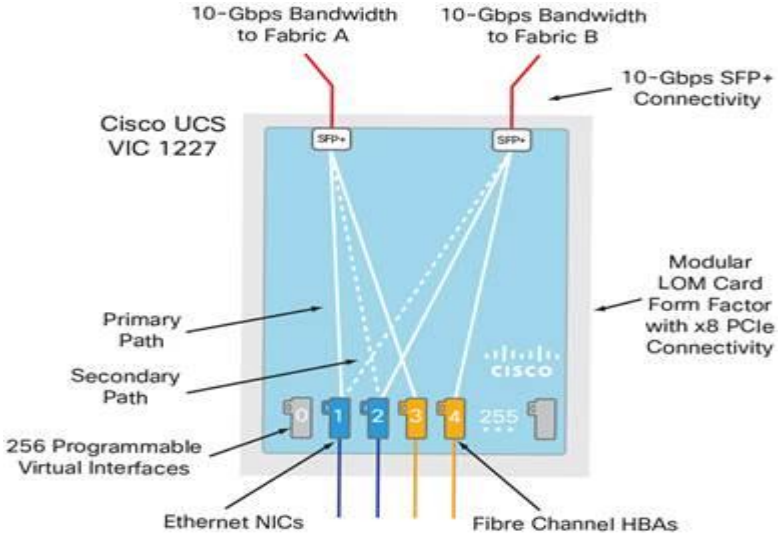
New to Cisco rack servers, the mLOM slot can be used to install a VIC without consuming a PCIe slot, providing greater I/O expandability. The VIC 1227 incorporates next-generation converged network adapter (CNA) technology from Cisco, providing investment protection for future feature releases. The card enables a policy-based, stateless, agile server infrastructure that can present up to 256 PCIe standards-compliant interfaces to the host, which can be dynamically configured as either NICs or HBAs. In addition, the VIC 1227 supports Cisco® Data Center Virtual Machine Fabric Extender (VM-FEX) technology, which extends the Cisco UCS fabric interconnect ports to virtual machines, simplifying server virtualization deployment.

For more information about the VIC, see:

- http://www.cisco.com/c/en/us/products/interfaces-modules/ucs-virtual-interface-card-1227/index.html
- http://www.cisco.com/c/en/us/products/collateral/interfaces-modules/unified-computing-system-adapters/datasheet-c78-732515.html

Figure 5. Cisco UCS VIC 1227 CNA

Figure 6. Cisco UCS VIC 1227 CNA Architecture



NVIDIA GRID Cards

For desktop virtualization applications, the GRID K1 and K2 cards are an optimal choice for high graphics performance (Table 2).

Table 2. Technical Specification for NVIDIA GRID Cards

	NVIDIA GRID K1 	NVIDIA GRID K2 
GPU	4 Kepler GK107	2 High End Kepler GK104
CUDA cores	768 (192/GPU)	3072 (1536/GPU)
Memory Size	16GB DDR3 (4GB/GPU)	8GB GDDR5
Max Power	130 W	225 W
Aux power requirement	6-pin connector	8-pin connector
PCIe	X16	X16
OpenGL	4.x	4.x
Microsoft DirectX	11	11
vGPU support	Yes	Yes
# users	4 –100[1]	2 – 64[1]

NVIDIA GRID Technology

The NVIDIA GRID virtualization solution is built on more than 20 years of software and hardware innovations in the accelerated graphics field to deliver a rich graphics experience to users running virtual desktops or applications.

For more information about NVIDIA GRID technology, see http://www.nvidia.com/object/grid-technology.html.

GRID GPUs

NVIDIA's Kepler-based GRID K1 and K2 boards are specifically designed to enable rich graphics in virtualized environments. They offer these main features:

- **High user density:** GRID boards have an optimized multiple-GPU design that helps increase user density. The GRID K1 board has four GPUs and 16 GB of graphics memory. In combination with NVIDIA GRID vGPU technology, the GRID K1 supports up to 32 users on a single board.
- **Power efficiency:** GRID boards are designed to provide data center–class power efficiency, including the revolutionary new streaming multiprocessor, SMX. The result is an innovative, proven solution that delivers revolutionary performance per watt (W) for the enterprise data center.
- **Reliability 24 hours a day, 7 days a week:** GRID boards are designed, built, and tested by NVIDIA for operation all day, every day. Working closely with Cisco helps ensure that GRID cards perform optimally and reliably for the life of the system.

For more information about GRID boards, see http://www.nvidia.com/object/grid-technology.html.

NVIDIA GRID Accelerated Remoting

Low-Latency Remote Display
NVIDIA's patented low-latency remote display technology greatly improves the user experience by reducing the lag that users feel when interacting with a virtual machine. With this technology, the virtual desktop screen is encoded and pushed directly to the remoting protocol.

Built into every Kepler GPU is a high-performance H.264 encoding engine capable of encoding simultaneous streams with superior quality. This feature greatly enhances cloud server efficiency by offloading encoding functions from the CPU and allowing the encoding function to scale with the number of GPUs in a server.

The GRID Accelerated Remoting technology is available in industry-leading remote display protocols such as VMware Horizon View and NICE DCV.

GRID Virtualization

Widest Range of Virtualization Solutions
GRID cards enable GPU-capable virtualization solutions from Microsoft and VMware, delivering the flexibility to choose from a wide range of proven solutions. GRID GPUs can be dedicated to a single high-end user or shared among multiple users.

NVIDIA GRID vGPU

GRID boards feature NVIDIA Kepler-based GPUs that, for the first time, allow hardware virtualization of the GPU. Thus, multiple users can share a single GPU, improving user density while providing true PC performance and application compatibility.

For more information about GRID vGPU, see http://www.nvidia.com/object/virtual-gpus.html.

VMware vSphere 6.0

VMware provides virtualization software. VMware's enterprise software hypervisors for servers—VMware vSphere ESX, vSphere ESXi, and VSphere—are bare-metal hypervisors that run directly on server hardware without requiring an additional underlying operating system. VMware vCenter Server for vSphere provides central management and complete control and visibility into clusters, hosts, virtual machines, storage, networking, and other critical elements of your virtual infrastructure.

vSphere 6.0 introduces many enhancements to vSphere Hypervisor, VMware virtual machines, vCenter Server, virtual storage, and virtual networking, further extending the core capabilities of the vSphere platform.

VMware ESXi 6.0 Hypervisor

What's New in the VMware vSphere 6.0 Platform

Scalability Improvements

ESXi 6.0 dramatically increases the scalability of the platform. With vSphere Hypervisor 6.0, clusters can scale to as many as 64 hosts, up from 32 in previous releases. With 64 hosts in a cluster, vSphere 6.0 can support 8000 virtual machines in a single cluster. This capability enables greater consolidation ratios, more efficient use of VMware vSphere Distributed Resource Scheduler (DRS), and fewer clusters that must be separately managed. Each vSphere Hypervisor 6.0 instance can support up to 480 logical CPUs, 12 terabytes (TB) of RAM, and 1024

virtual machines. By using the newest hardware advances, ESXi 6.0 enables the virtualization of applications that previously had been thought to be nonvirtualizable.

ESXi Security Enhancements

ESXi 6.0 offers these security enhancements:

- **Account management:** ESXi 6.0 enables management of local accounts on the ESXi server using new ESXi CLI commands. The capability to add, list, remove, and modify accounts across all hosts in a cluster can be centrally managed using a vCenter Server system. Previously, the account and permission management functions for ESXi hosts were available only for direct host connections. The setup, removal, and listing of local permissions on ESXi servers can also be centrally managed.
- **Account lockout:** ESXi Host Advanced System Settings have two new options for the management of failed local account login attempts and account lockout duration. These parameters affect Secure Shell (SSH) and vSphere Web Services connections, but not ESXi Direct Console User Interface (DCUI) or console shell access.
- **Password complexity rules:** In previous versions of ESXi, password complexity changes had to be made by manually editing the /etc/pam.d/passwd file on each ESXi host. In vSphere 6.0, an entry in Host Advanced System Settings enables setting changes to be centrally managed for all hosts in a cluster.
- **Improved auditability of ESXi administrator actions:** Prior to vSphere 6.0, actions at the vCenter Server level by a named user appeared in ESXi logs with the vpxuser username: for example, [user=vpxuser]. In vSphere 6.0, all actions at the vCenter Server level for an ESXi server appear in the ESXi logs with the vCenter Server username: for example, [user=vpxuser: DOMAIN\User]. This approach provides a better audit trail for actions run on a vCenter Server instance that conducted corresponding tasks on the ESXi hosts.
- **Flexible lockdown modes:** Prior to vSphere 6.0, only one lockdown mode was available. Feedback from customers indicated that this lockdown mode was inflexible in some use cases. With vSphere 6.0, two lockdown modes are available:
  - In normal lockdown mode, DCUI access is not stopped, and users on the DCUI.Access list can access the DCUI.
  - In strict lockdown mode, the DCUI is stopped.

In addition, vSphere 6.0 offers a new function called exception users. Exception users are local accounts or Microsoft Active Directory accounts with permissions defined locally on the host to which these users have host access. These exception users are not recommended for general user accounts but are recommended for use by third-party applications—for service accounts, for example—that need host access when either normal or strict lockdown mode is enabled. Permissions on these accounts should be set to the bare minimum required for the application to perform its task and with an account that needs only read-only permissions on the ESXi host.

- **Smart card authentication to DCUI:** This function is for U.S. federal customers only. It enables DCUI login access using a Common Access Card (CAC) and Personal Identity Verification (PIV). The ESXi host must be part of an Active Directory domain.

Graphics Acceleration in VMware Horizon View

Virtual desktop infrastructure (VDI) solutions are becoming increasingly common in standard office environments. Today, bulky desktop systems can be replaced with small thin clients, delivering Microsoft Windows desktops from the data center and enabling a more flexible way of working in multiple locations or from home. But this new model presents challenges for users who typically require a special workstation with high-end graphic cards and a dual-monitor setup and who use graphics-intensive applications. GRID technology addresses these challenges.

Graphics Acceleration in VMware Horizon View 6.1 with GRID vGPU

Using GRID vGPU technology, the graphics commands of each virtual machine are passed directly to the GPU, without translation by the hypervisor. This feature enables the GPU hardware to be time sliced, to deliver excellent shared virtualized graphics performance. The GRID vGPU offers greater flexibility than any other solution, enabling deployment of virtual machines across a wide range of users and graphics applications, including Microsoft PowerPoint slides and YouTube videos and the most demanding 3D computer-aided design (CAD) software (Table 3).

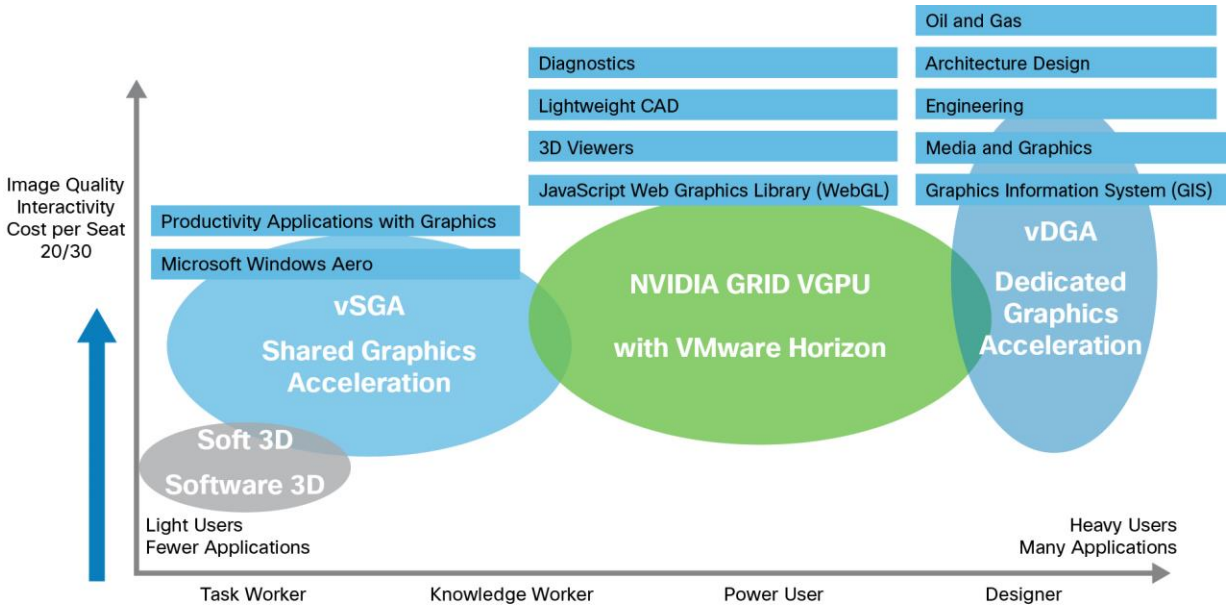
3D graphics capabilities in Horizon View further expand the target user base and potential use cases that IT can accommodate with virtual desktops. In addition, 3D augments the virtual desktop user interface by enabling a more graphically rich user experience.

Table 3. GPU Deployment and Supported Applications

Attributes	vSGA	NVIDIA GRID VGPU	vDGA
Sharing	GPUs shared between users	GPUs shared between users	**GPU dedicated to one user**
Consolidation	Good consolidation for low-end graphics use cases	Good consolidation ratio (up to 8:1)	**Only one user per GPU (1:1)**
Performance	**Lightweight rich graphics WITHOUT video acceleration**	Scalable performance from entry level to high end	High end workstation performance
App Compatibility	**Drivers NOT application certified**	Fully certified application drivers	Fully certified application drivers
DirectX APIs	**DirectX 9**	DirectX 9, 10, 11	DirectX 9, 10, 11
OpenGLAPIs	**OpenGL 2.1**	OpenGL 2.1, 3.x, 4.x	OpenGL 2.1, 3.x, 4.x
General Purpose Compute	**Does NOT support compute CUDA, OpenCL**	**Does NOT support compute CUDA, OpenCL**	Compute APIs with CUDA, OpenCL
Automated Management	Yes	Yes	**No**
vMotion/HA	Yes	**No**	No

With the support of Horizon and the GRID vGPU, GPU deployment is becoming an increasingly popular and cost-effective way to meet user and application requirements (Figure 7).

Figure 7. Virtual Desktop User Segmentation and Application



Difference Between Soft 3D, vSGA, vDGA, and vGPU

Table 4 summarizes the differences between Soft 3D, vSGA, vDGA, and vGPU graphics drivers.

Table 4. Graphics Driver Comparison

Name	Definition	Description
Soft 3D	Software 3D renderer	Support for software-accelerated 3D graphics is provided through a VMware Microsoft Windows Display Driver Model (WDDM) 1.1–compliant driver without the need to have any physical GPUs installed in the ESXi host.
vSGA	Virtual shared graphics acceleration	Multiple virtual machines use physical GPUs installed locally in the ESXi hosts to provide hardware-accelerated 3D graphics to multiple virtual desktops.
vDGA	Virtual dedicated graphics acceleration	Only one virtual machine is mapped to a single physical GPU installed in the ESXi host to provide high-end, hardware-accelerated workstation graphics when a discrete GPU is needed.
vGPU	Virtual graphics processing unit	Multiple virtual machines use physical GPUs installed locally in the ESXi hosts, and the virtual machines or users share the GPU resources through a shared PCI mode profile attached to the user to provide shared graphics.

Soft 3D: Software-Based 3D Rendering

The Soft 3D renderer is based on a VMware WDDM 1.1–compliant driver and is installed with VMware Tools on Microsoft Windows 7 virtual desktops. Soft 3D differs from vSGA and vDGA in that it does not require any physical GPUs to be installed in the ESXi host.

The VMware Soft 3D graphics driver provides support for DirectX 9.0c and OpenGL 2.1. The driver is supported on Windows 7 for 2D and 3D graphics and is used for both Soft 3D and vSGA. vDGA configurations do not use the VMware Soft 3D driver; instead, they use the native graphics-card driver installed directly in the guest OS.

One of the benefits of VMware Soft 3D for both software 2D and 3D and vSGA implementations is that a virtual machine can dynamically switch between software and hardware acceleration without the need for reconfiguration. Additionally, sharing this driver allows the use of VMware high-availability technologies such as VMware vSphere vMotion. The use of a single driver also greatly simplifies image management and deployment.

Note: If you are dynamically moving from hardware 3D rendering to software 3D rendering, you may notice a performance drop in applications running in the virtual machine. However, if you are moving in the reverse direction (software to hardware), you should notice an improvement in performance.

vSGA: Virtual Shared Graphics Acceleration

To provide hardware-accelerated 3D graphics, vSGA allows multiple virtual machines to use physical GPUs installed locally in the ESXi hosts. This approach differs from Soft 3D in that physical GPUs must be installed in the host server. These GPUs are shared across multiple virtual machines, unlike vDGA, in which each virtual machine is directly mapped to a single GPU.

The maximum amount of video memory that can be assigned per virtual machine is 512 MB. However, video-memory allocation is evenly divided: Half the video memory is reserved on the hardware GPU, and the other half is reserved through host RAM. (You need to take this division into consideration when sizing ESXi host RAM.)

You can use this rule to calculate basic consolidation ratios. For example, the GRID K1 card has 16 GB of GPU RAM. If all virtual machines are configured with 512 MB of video memory, half of which (256 MB) is reserved on the GPU, you can calculate that a maximum of 64 virtual machines can run on that specific GPU at any given time.

The ESXi host reserves GPU hardware resources on a first-come, first-served basis as virtual machines are powered on. If all GPU hardware resources are already reserved, additional virtual machines will be unable to power on if they are explicitly set to use hardware 3D rendering. If the virtual machines are set to automatic mode, they will be powered on using software 3D rendering.

vSGA is limited by the amount of memory on the installed boards. ESXi assigns a virtual machine to a particular graphics device during power-on. The assignment is based on graphics-memory reservation that occurs in a round-robin fashion. The current policy is to reserve one-half of the virtual machine's VRAM size with a minimum of 128 MB. Therefore, a graphics device with a 4 GB of memory can accept at most 32 virtual machines with a minimum reservation. After a graphics device reaches its reservation maximum, no more virtual machines will be assigned to it until another virtual machine leaves the GPU. This change can occur when a virtual machine is powered off, suspended, or moved through vMotion to another host.

vDGA: Virtual Dedicated Graphics Acceleration

vDGA graphics-acceleration capability is provided by ESXi to deliver high-end workstation graphics for use cases in which a discrete GPU is needed. This graphics-acceleration method dedicates a single GPU to a single virtual machine for high performance.

Note: Some graphics cards can run multiple GPUs.

If you are using vDGA, graphics adapters installed in the underlying host are assigned to virtual machines using vSphere DirectPath I/O. Assigning a discrete GPU to the virtual machine dedicates the entire GPU to that virtual machine.

For vDGA, the number of 3D-enabled virtual machines is limited by the number of GPUs in the server. A C240 M4 can accommodate two cards. The GRID K2 card has two GPUs. If an administrator installs a GRID K2 card in each

available slot on the C240 M4, the system will have a total of four GPUs. This will be the total number of vDGA-enabled virtual machines that the server can support.

Note: Both vSGA and vDGA can support a maximum of eight GPU cards per ESXi host.

vGPU: Virtual Graphics Processing Unit

Horizon 6.1 and vSphere 6 with the GRID vGPU enable designers, architects, and engineers to run the most advanced, graphics-intensive applications on a remote desktop using NVIDIA professional 3D graphics and certified application drivers. vGPU brings workstation-class performance to remote and mobile workers more affordably than ever before, even over high-latency networks.

vGPU software, combined with GRID K1 and K2 graphics adapters, offers a platform that delivers true GPU hardware acceleration shared between multiple virtual desktops. vGPU sharing is accomplished with different GPU profiles that each enable a dedicated amount of video memory appropriate for different use cases, just as with different physical graphics adapters.

Software Requirements for vDGA, vSGA, and vGPU

Table 5 lists the software requirements for vDGA, vSGA, and vGPU rendering.

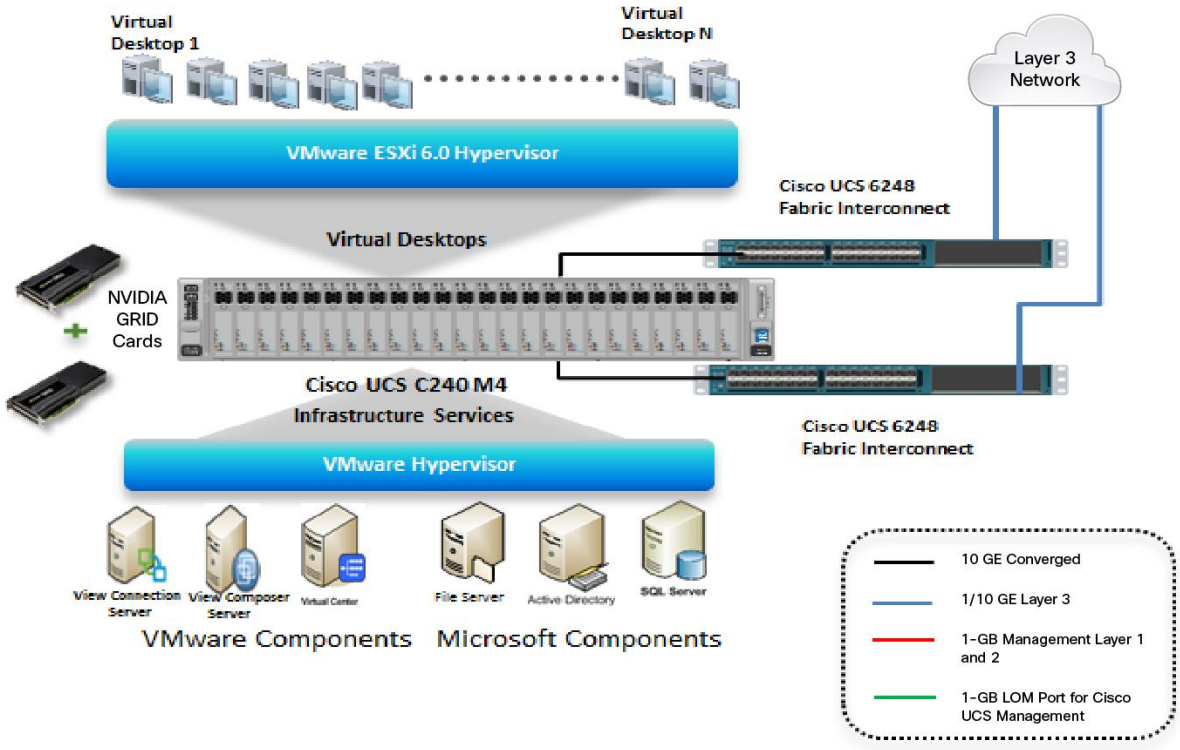
Table 5. vDGA, vSGA, and vGPU Software Requirements

Product	Description
VMware vSphere Hypervisor	• vSGA and vDGA: ESXi 5.1 U1 or ESXi 5.5 (ESXi 5.5 recommended) • vGPU: ESXi 6.1 and vSphere 6.0
VMware Horizon View	• vSGA: Horizon View 5.0 or later (Horizon View 5.3 recommended) • vDGA: Horizon View 5.0 or later • vGPU: Horizon View 6.1 or later
Display protocol	vSGA, vDGA, and vGPU: PC over IP (PCoIP) with a maximum of 4 display monitors
NVIDIA drivers	• vGPU: NVIDIA drivers for vSphere ESXi 6.0 Version 346.68/348.07 (http://www.nvidia.com/download/driverResults.aspx/85390/en-us) • vSGA: NVIDIA drivers for vSphere ESXi 6.0 Version 346.69 (http://www.nvidia.com/download/driverResults.aspx/85391/en-us) • vDGA: Tesla or GRID desktop driver Version 348.07 (http://www.nvidia.com/download/driverResults.aspx/86814/en-us) **Note:** vDGA virtual machine drivers are compatible with Windows 7 64-bit, Windows 8.1 64-bit, Windows 8 64-bit, and Windows Vista 64-bit. **Note:** These drivers are supplied and supported by NVIDIA. Both drivers can be downloaded from http://www.nvidia.com/Download/index.aspx?lang=en-us.
Guest operating system	vSGA: Windows 7 32-bit or 64-bit vGPU and vDGA: Windows 7 64-bit

Solution Configuration

Figure 8 provides an overview of the solution configuration.

Figure 8. Reference Architecture



The hardware components of the solution are:

- Cisco UCS C240-M4 Rack Server (2 Intel Xeon processor E5-2680 v3 CPUs at 2.50 GHz) with 384 GB of memory (16 GB x 24 DIMMs at 2133 MHz), and hypervisor host
- Cisco UCS VIC1227 mLOM
- 2 Cisco Nexus® 5548UP Switches (access switches)
- 2 Cisco UCS 6248UP 48-Port Fabric Interconnects
- 12 x 600-GB SAS disks at 10,000 rpm
- NVIDIA GRID K1 and K2 cards

The software components of the solution are:

- Cisco UCS firmware 2.2(4b)
- VMware ESXi 6.0 for VDI hosts
- VMware Horizon View 6.1
- Microsoft Windows 7 SP1 32-bit, 2 virtual CPUs, and 2 GB of memory (for vSGA)
- Microsoft Windows 7 SP1 64-bit, 4 virtual CPUs, and 8 GB of memory (for vDGA)
- Microsoft Windows 7 SP1 64-bit, 4 virtual CPUs, and 8 GB of memory (for vGPU)

Configure Cisco UCS

This section describes the Cisco UCS configuration.

Install NVIDIA GRID or Tesla GPU Card on Cisco UCS C240 M4

Install the GPU card on the C240 M4 server. Table 6 lists the minimum firmware required for the GPU cards.

Table 6. Minimum Server Firmware Versions Required for GPU Cards

Cisco Integrated Management Controller	BIOS Minimum Version
NVIDIA GRID K1	2.0(3a)
NVIDIA GRID K2	2.0(3a)
NVIDIA Tesla K10	2.0(3e)
NVIDIA Tesla K20	2.0(3e)
NVIDIA Tesla K20X	2.0(3e)
NVIDIA Tesla K40	2.0(3a)

For more information, see http://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/c/hw/C240M4/install/C240M4/replace.html - pgfId-1372776.

Note the following NVIDIA GPU card configuration rules:

- You can mix GRID K1 and K2 GPU cards in the same server.
- Do not mix GRID GPU cards with Tesla GPU cards in the same server.
- Do not mix different models of Tesla GPU cards in the same server.
- All GPU cards require two CPUs and at least two 1400W power supplies in the server.

For more information, see:

- http://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/c/hw/C240M4/install/C240M4/replace.html - pgfId-1372848
- http://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/c/hw/C240M4/install/C240M4.pdf

The rules for configuring the server with GPUs differ, depending on the server version and other factors. Table 7 lists rules for populating the C240 M4 with NVIDIA GPUs. Figure 9 shows a one-GPU installation, and Figure 10 shows a two-GPU installation.

Table 7. NVIDIA GPU Population Rules for Cisco UCS C240 M4 Rack Server

Single GPU	Dual GPU
Riser 1A, slot 2 or Riser 2, slot 5	Riser 1A, slot 2 and Riser 2, slot 5

Note: When you install a GPU card in slot 2, Network Communications Services Interface (NCSI) support in riser 1 automatically moves to slot 1. When you install a GPU card in slot 5, NCSI support in riser 2 automatically moves to slot 4. Therefore, you can install a GPU card and a Cisco UCS VIC in the same riser.

Figure 9. One-GPU Scenario

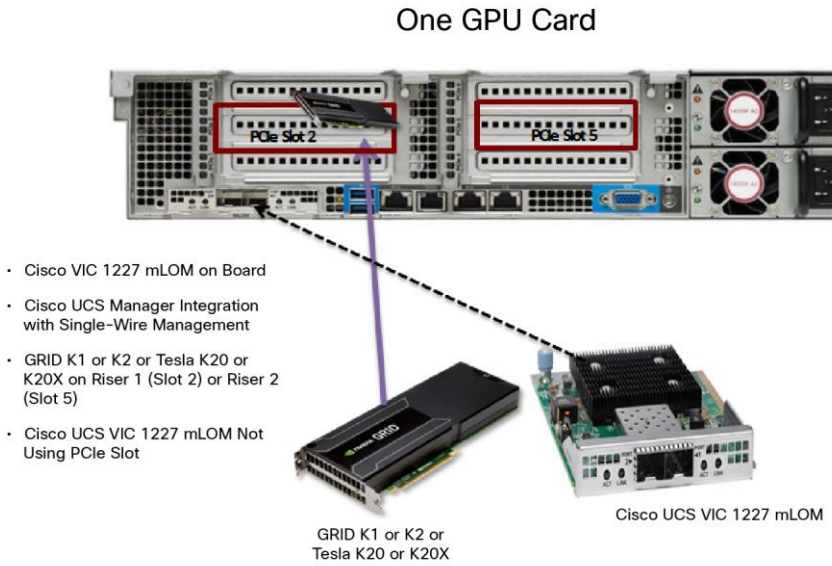
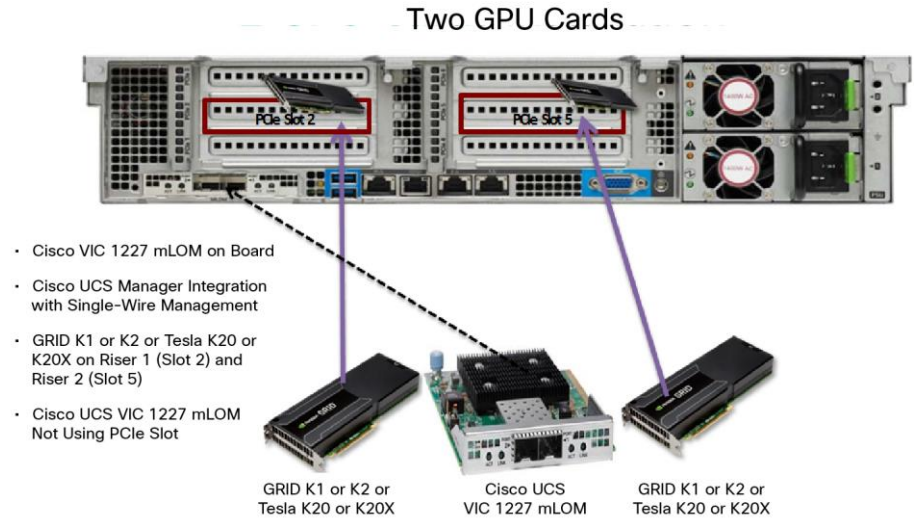


Figure 10. Two-GPU Scenario



For information about the physical configuration of GRID cards in riser slots 2 and 5 and for GPU card PCIe slot and support information, see: http://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/c/hw/C240M4/install/C240M4/replace.html - 75263.

Specify Base Cisco UCS Configuration

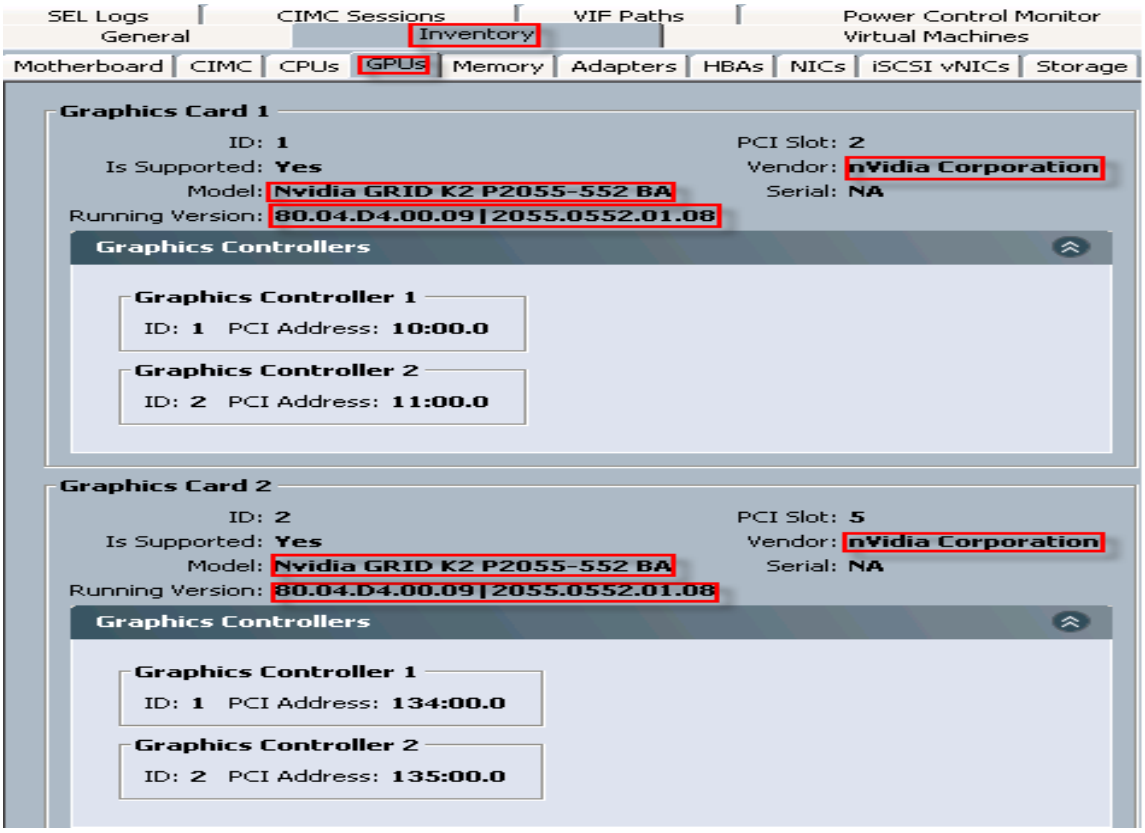
To configure physical connectivity and implement best practices for C-Series server integration with Cisco UCS Manager, see http://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/release/notes/ucs_2_2_rn.html.

Configure the GPU Card

1. After the server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs.

As shown in Figure 11, PCIe slots 2 and 5 are used with two GRID K2 cards running firmware Version 80.04.D4.00.09 | 2055.0552.01.08.

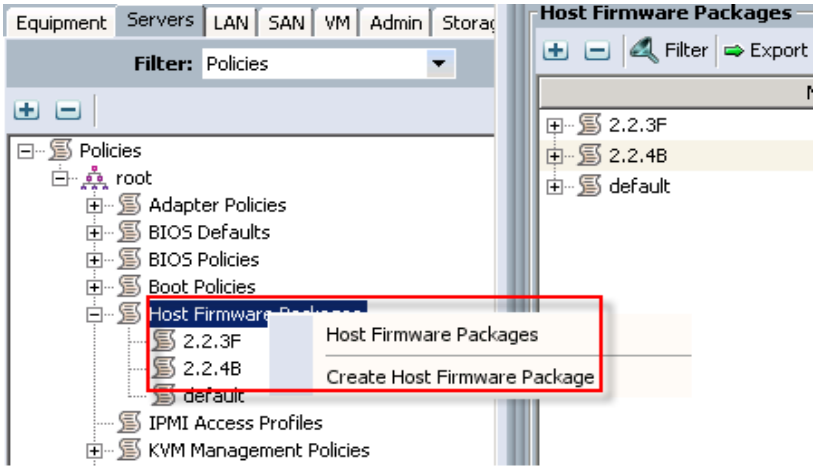
Figure 11. FiguNVIDIA GRID Cards Inventory Display on Cisco UCS Manager



You also can flash the memory cards on Cisco UCS manager and perform firmware upgrades.

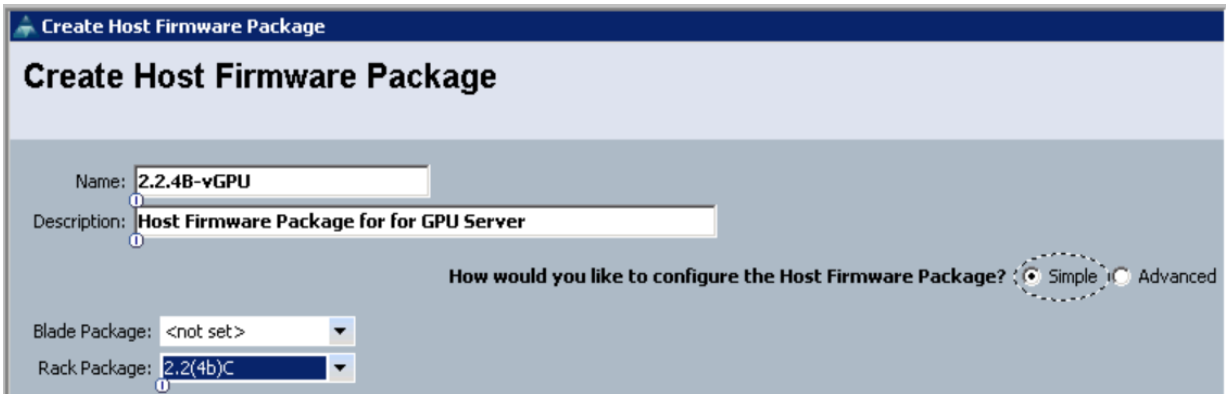
2. Create host firmware policy by selecting the Servers tab on Cisco UCS Manager. Choose Policies > Host Firmware Packages. Right-click and select Create Host Firmware Package (Figure 12).

Figure 12. Cisco UCS Manager Firmware Package



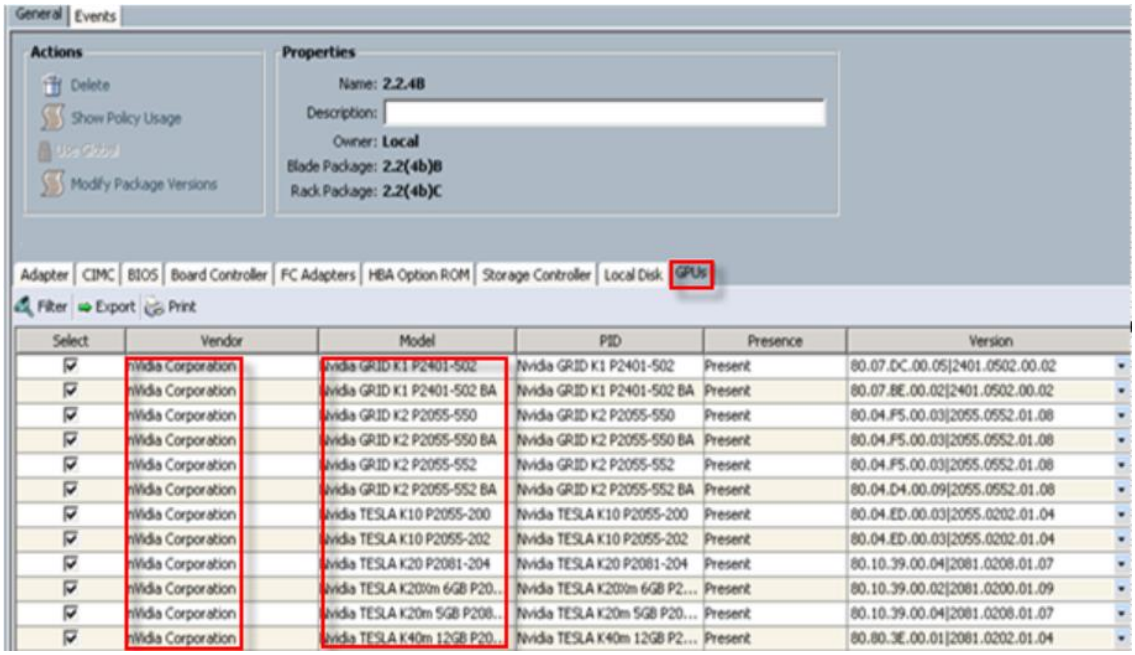
3. Select the Simple configuration of the host firmware package, and choose Rack Package 2.2(4b)C (Figure 13).

Figure 13. Cisco UCS Manager Firmware Package Configuration



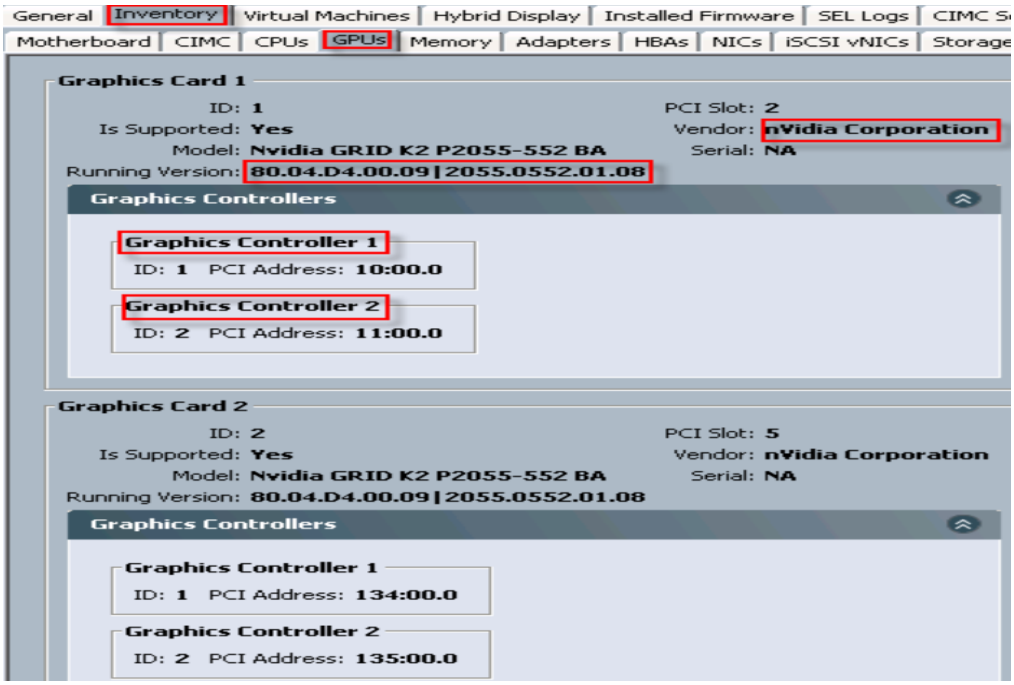
4. Click OK to display the list of firmware packages (Figure 14).

Figure 14. Cisco UCS Manager Firmware Package Selection for NVIDIA Components



5. Apply this host firmware package in the service profile template service profiles firmware policy. After the firmware upgrades are completed, the running firmware version for the GPUs is selected (Figure 15).

Figure 15. Running Firmware Version for the GPUs

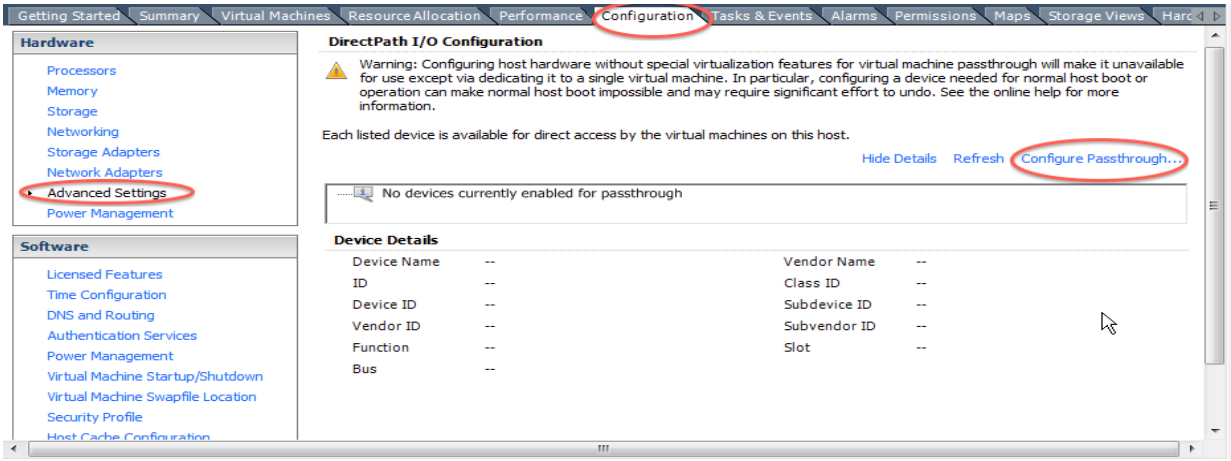


Note: Virtual machine hardware Version 9 or later is required for vSGA and vDGA configuration. Virtual machines with hardware Version 9 or later can have their settings managed only through the vSphere Web Client.

Configure vDGA or Pass-Through GPU Deployment

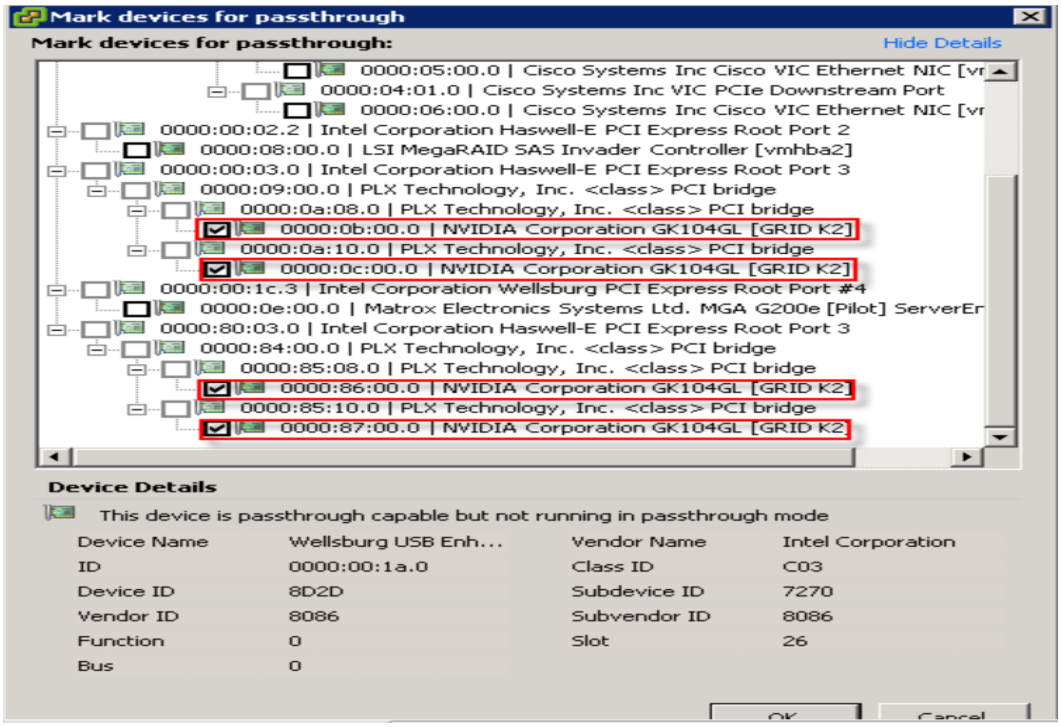
1. Select the ESXi host, choose Configuration, select the Hardware tab, and choose Advanced Settings > Configure Passthrough (Figure 16).

Figure 16. Enabling GRID Card Pass-Through Mode on ESXi Server



A dialog box will appear showing all the devices along with the GRID cards (Figure 17).

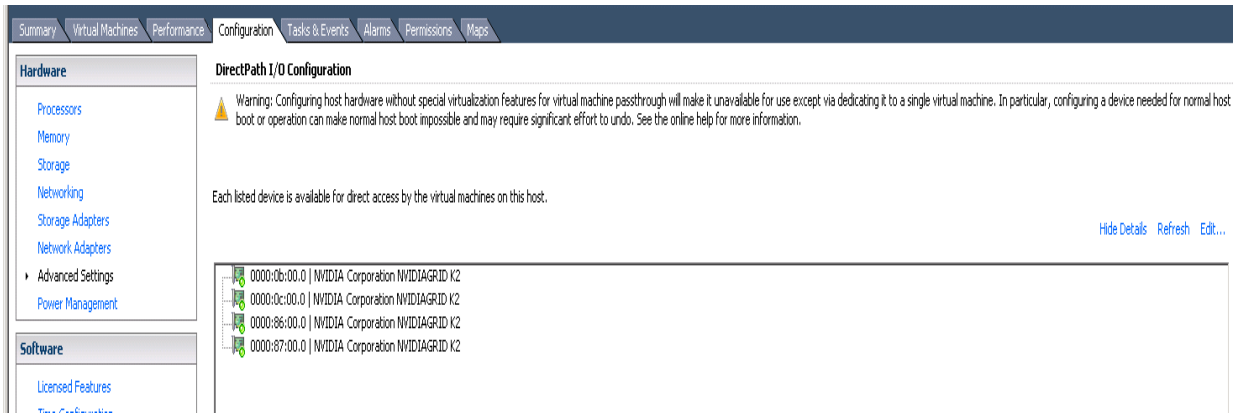
Figure 17. NVIDIA Pass-Through Devices Displayed on the ESXi Host



2. Select the NVIDIA GRID card types installed on the server from among the GRID card adapters.
3. After making changes for pass-through configuration, reboot the ESXi host (Figure 18).

Devices enabled for pass-through configuration for dedicated graphics deployment will be marked in green after you reboot. If a device isn't marked in green, use the VMware documentation to perform troubleshooting.

Figure 18. Displaying All Selected NVIDIA GRID GPU Card Adapters

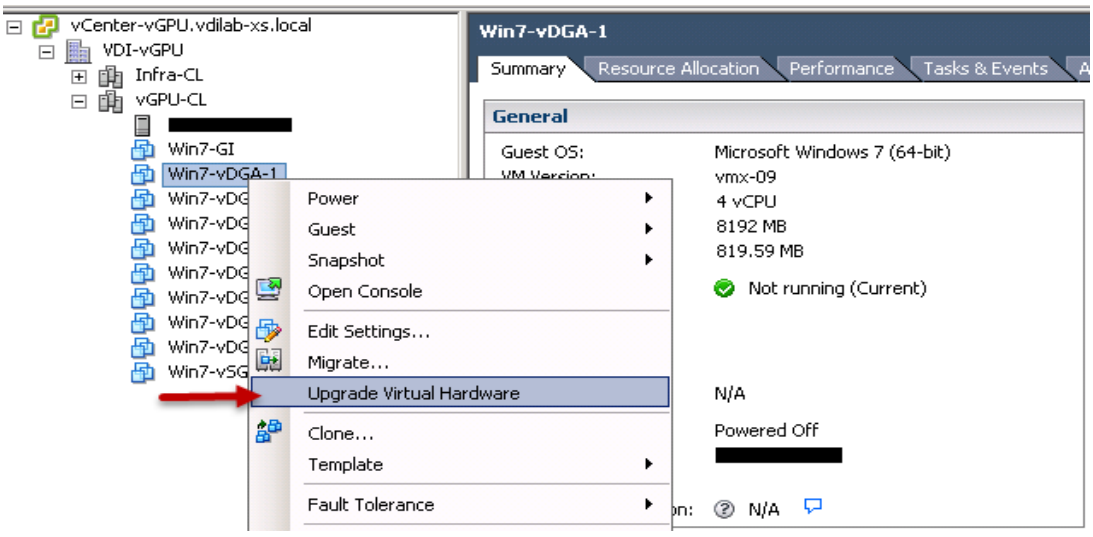


Note: vDGA does not support live vSphere vMotion capabilities. Bypassing the virtualization layer, vDGA uses vSphere DirectPath I/O to allow direct access to the GPU card. By enabling direct pass-through from the virtual machine to the PCI device installed on the host, you effectively lock the virtual machine to that specific host. If you need to move a vDGA-enabled virtual machine to a different host, you should power off the virtual machine, use vMotion to migrate it to another host that has a GPU card installed, and reenable pass-through to the specific PCI device on that host. Only then should you power on the virtual machine.

Enable Virtual Machine for vDGA Configuration

1. Upgrade the virtual machine hardware version. In the virtual machine Edit Settings options, choose Upgrade Virtual Hardware (Figure 19).

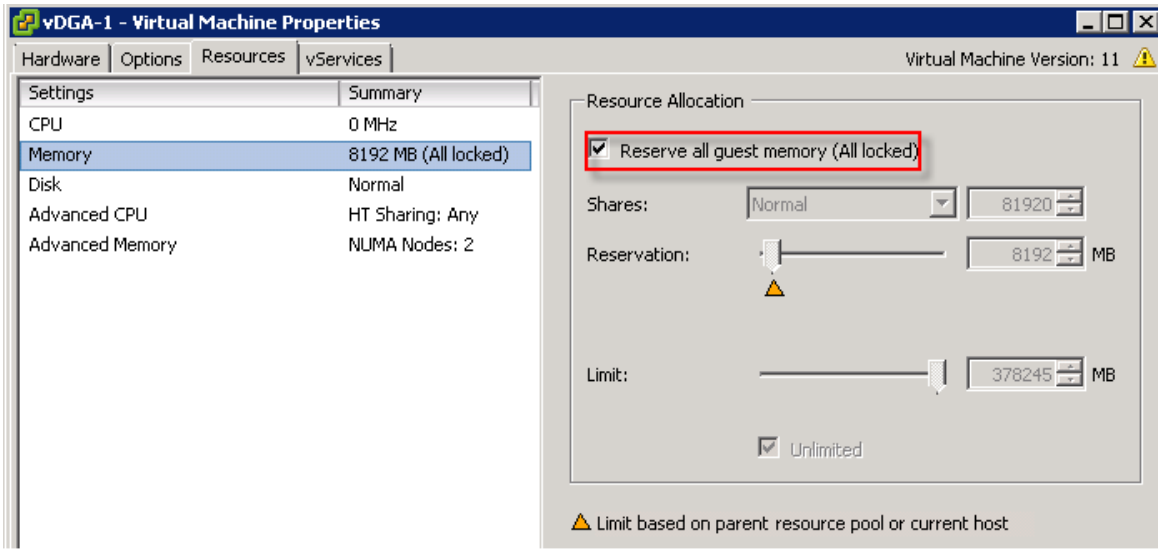
Figure 19. Upgrading the Virtual Machine



Note: If you are using virtual machine hardware Version 9.0 or earlier, you need to upgrade the virtual machine hardware to Version 11.

2. Reserve all guest memory. In the virtual machine Edit Settings options, select the Resources tab. Select "Reserve all guest memory (All locked)" as shown in Figure 20.

Figure 20. Reserving All Guest Memeory



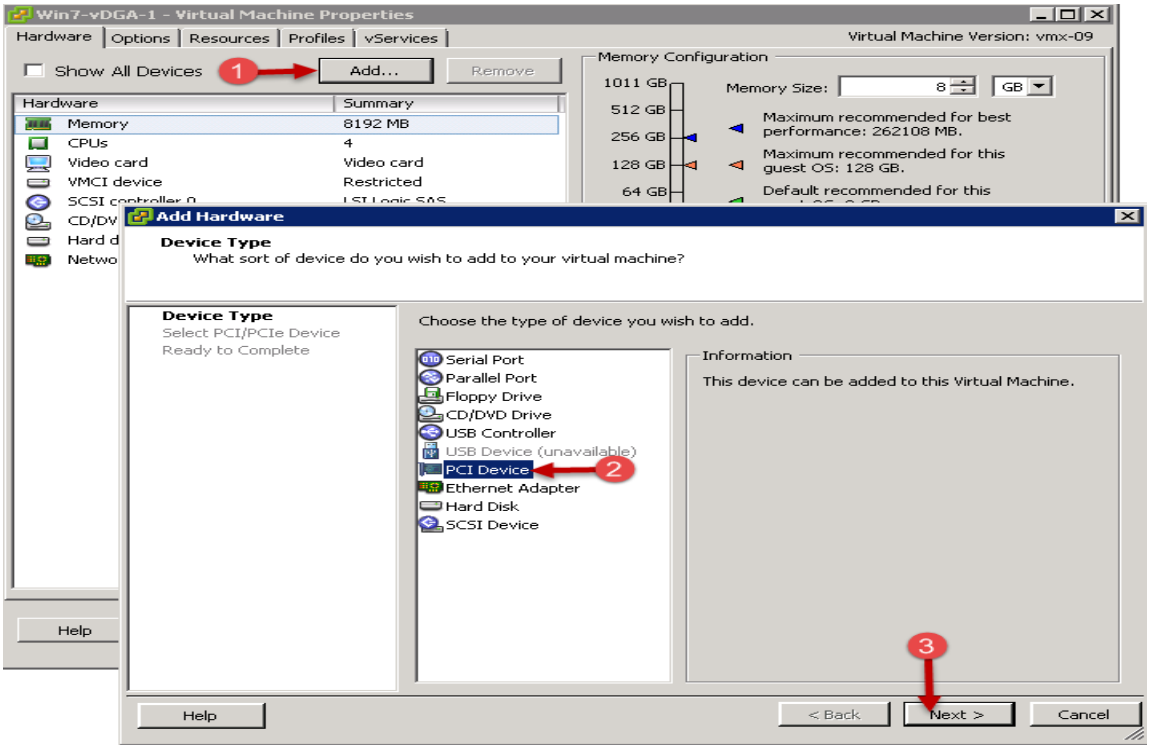
3. If the virtual machine has more than 2 GB of configured memory, adjust pciHole.start. Add the following parameter to the .vmx file of the virtual machine (you can add this parameter at the end of the file):
   pciHole.start = "2048"

Note: This step is required only if the virtual machine has more than 2 GB of configured memory.

4. Add a PCI device:

   a. In the virtual machine Edit Settings, select the Hardware tab. Click Add and select PCI Device. Click Next (Figure 21).

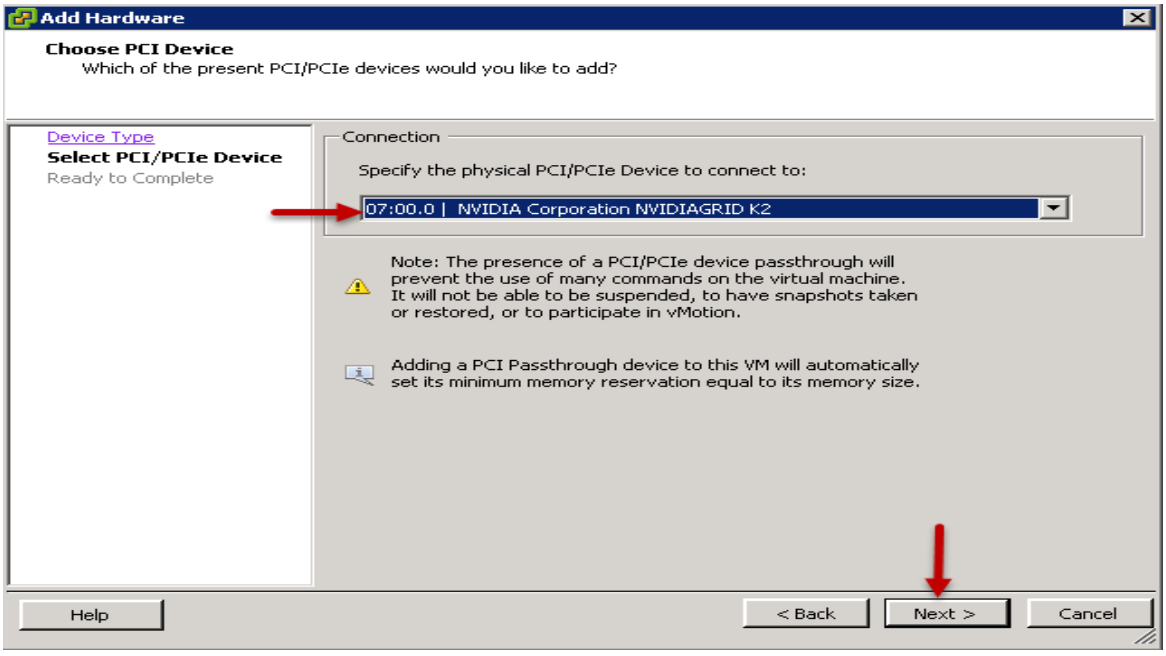
Figure 21. Adding a PCI Device to the Virtual Machine to Attach GPU Controller Pass-Through Mode



   b. Select the PCI device from the drop-down list. Click Next (Figure 22).

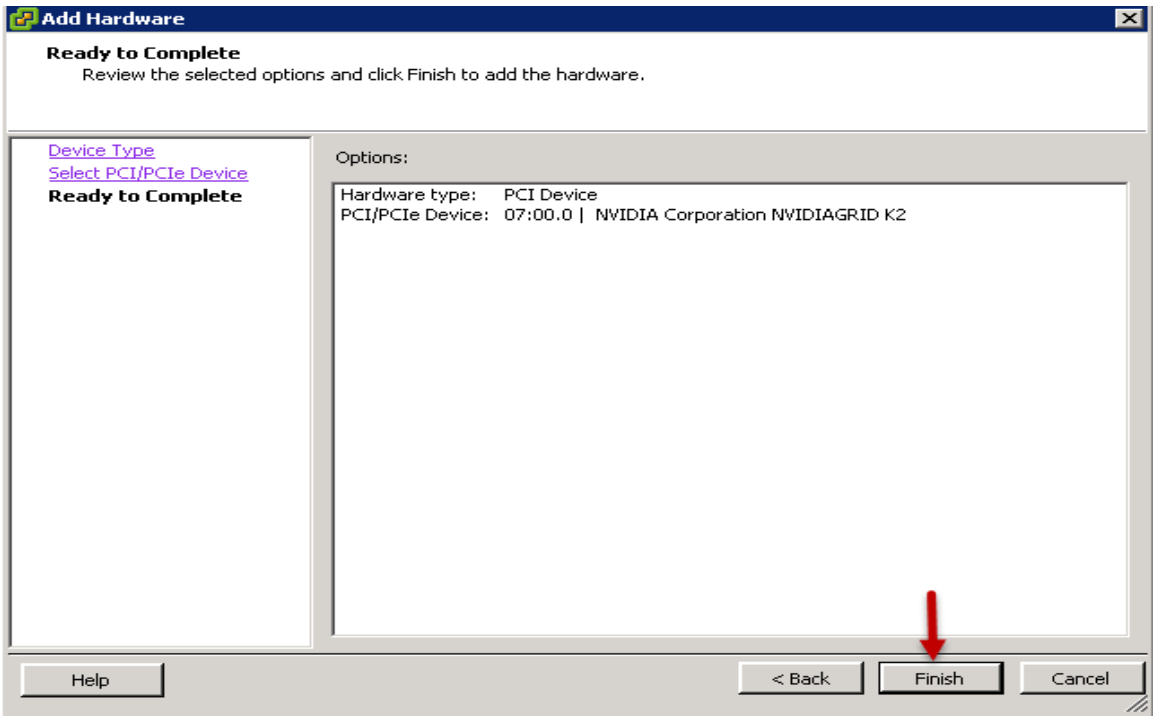
Note: Only one virtual machine can be powered on if the same PCI device is added to multiple virtual machines.

Figure 22. Selecting the NVIDIA GPU Controller for the Virtual Machine



c. Click Finish on the Ready to Complete screen (Figure 23).

Figure 23. Click Finish to Attach the Selected NVIDIA Controller to the Virtual Machine



5. Install the latest NVIDIA Windows 7 or 8 desktop drivers on the virtual machine. All NVIDIA drivers can be found at http://www.nvidia.com/Download/index.aspx?lang=en-us.

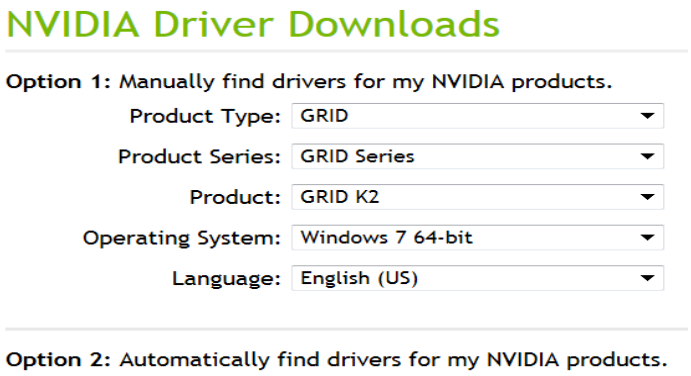
Note: Before installing NVIDIA drivers, optimize Microsoft Windows using VMware best practices.

Download Virtual Machine Drivers from NVIDIA Website

1. Download the virtual machine drivers from the NVIDIA website: http://www.nvidia.com/download/driverResults.aspx/86066/en-us (Figure 24).

Note: Select 32-bit or 64-bit graphics drivers based on the guest OS type.

Figure 24. Download NVIDIA Drivers



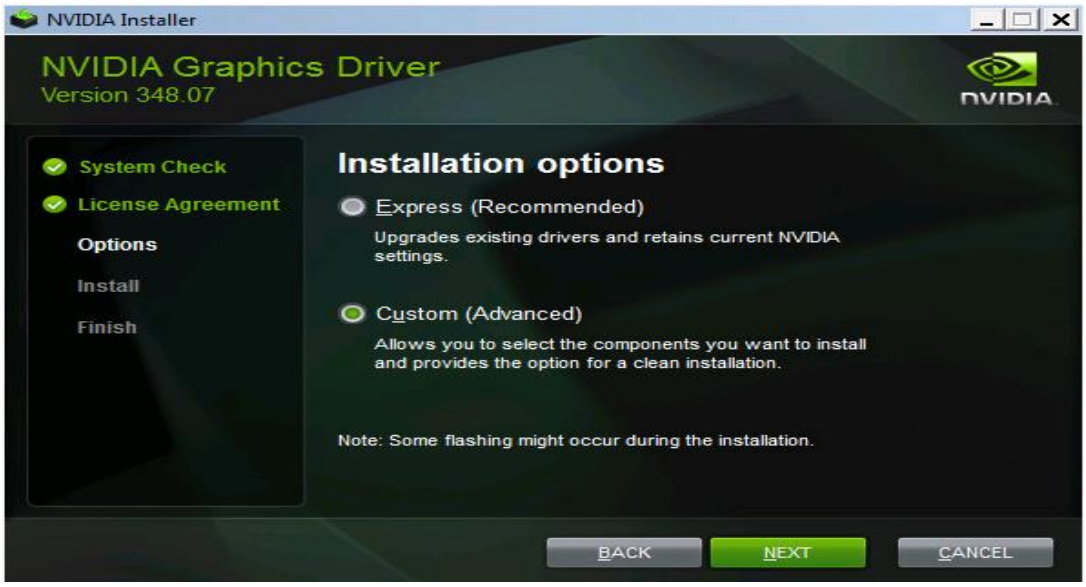
For more information see the VMware Horizon with View Optimization Guide for Windows 7 and Windows 8 Whitepaper: http://www.vmware.com/files/pdf/VMware-View-OptimizationGuideWindows7-EN.pdf

2. Install the drivers.
   a. Accept the license terms and agreement (Figure 25); then click Next.

Figure 25. NVIDIA Graphics Drivers installation system check

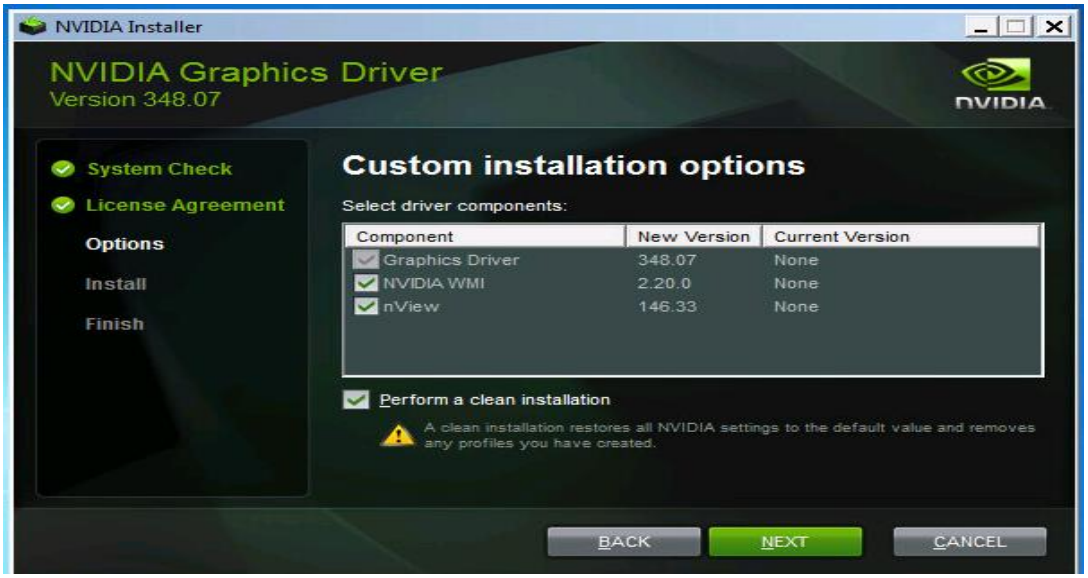
b. Select Custom (Advanced) installation. Click Next (Figure 26).

Figure 26. NVIDIA Graphics Driver Installation Options



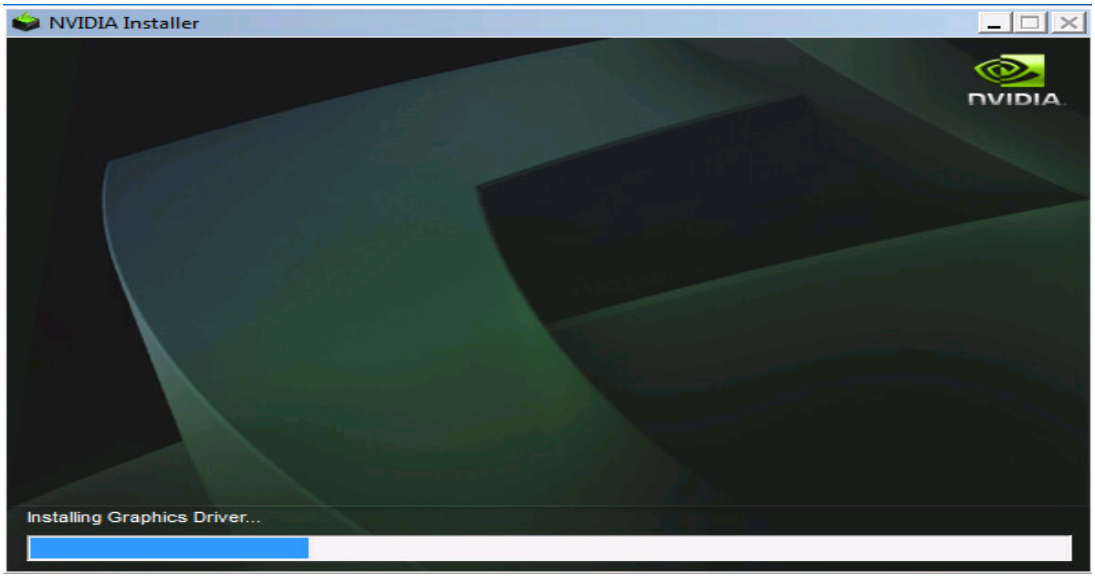
c. Select the check box for each option and select “Perform a clean installation.” Click Next (Figure 27).

Figure 27. Select All Components Available and Perform a Clean Installation



The drivers are installed (Figure 28).

Figure 28. Installation Progress



c. Click Finish. Reboot the virtual machine.

3. Install VMware View Agent. Reboot when prompted.
4. Enable the proprietary NVIDIA capture APIs.

   a. After the virtual machine has rebooted, enable the proprietary NVIDIA capture APIs by running C:\Program Files\Common Files\VMware\Teradici PCoIP Server\ MontereyEnable.exe and specifying –enable.

   b. After the process is complete, restart the virtual machine.

5. Verify that the virtual machine is using the NVIDIA GPU and driver.

To activate the NVIDIA display adapter, you must connect to the virtual machine for the first time through PCoIP in full-screen mode from the endpoint (at native resolution), or else the virtual machine will use the Soft 3D display adapter. vDGA does not work through the vSphere console.

After the virtual machine has rebooted and you have connected through PCoIP in full-screen mode, verify that the GPU is active by viewing the display information in DXDiag.exe:

a. Click the Start menu from the virtual machine to which the GRID card pass-through device is attached.

b. Type **dxdiag** and when DxDiag appears in the list, click Enter; or click DxDiag in the list.

c. After DxDiag launches, click the Display tab to verify that the virtual machine is using the NVIDIA GPU and driver (Figure 29).

Figure 29. Launching DxDiag from the Virtual Machine to Check the Driver Status

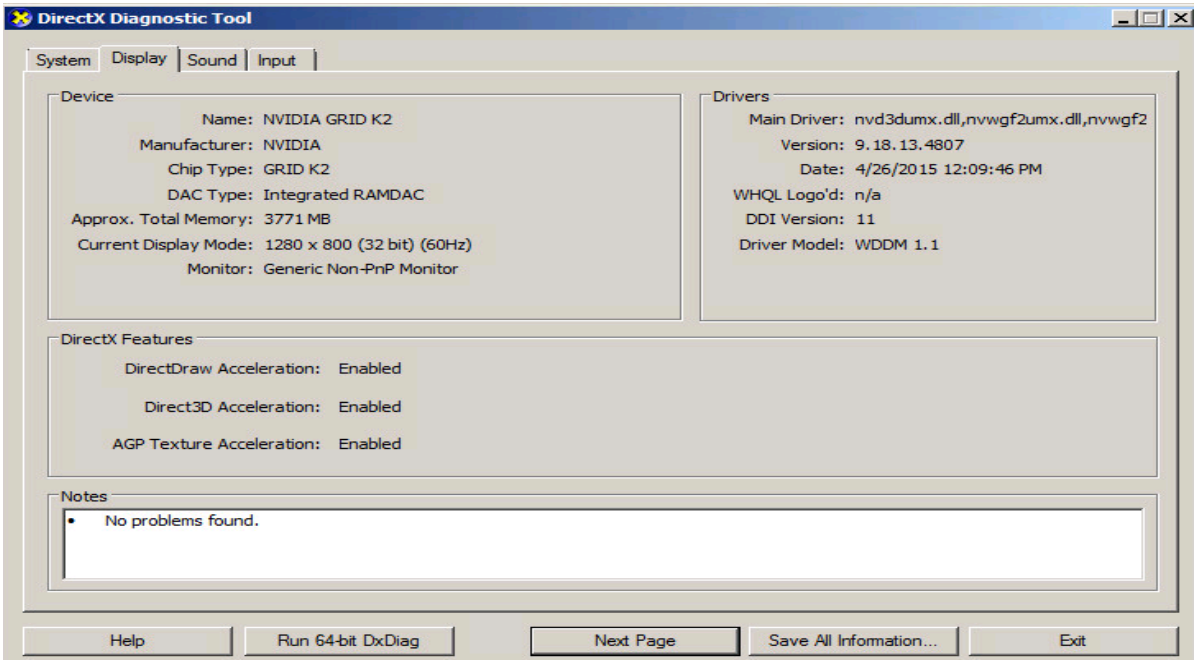
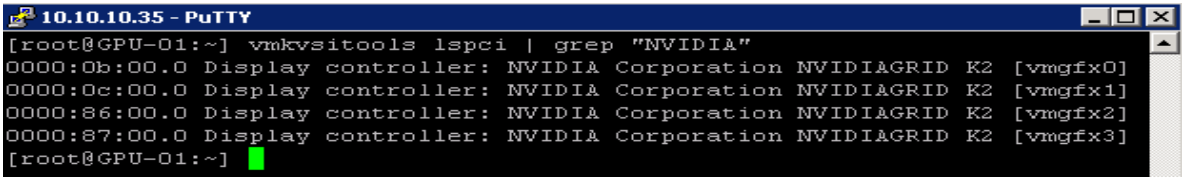


Figure 30. Displaying All the GRID Card Controllers Present on the Host

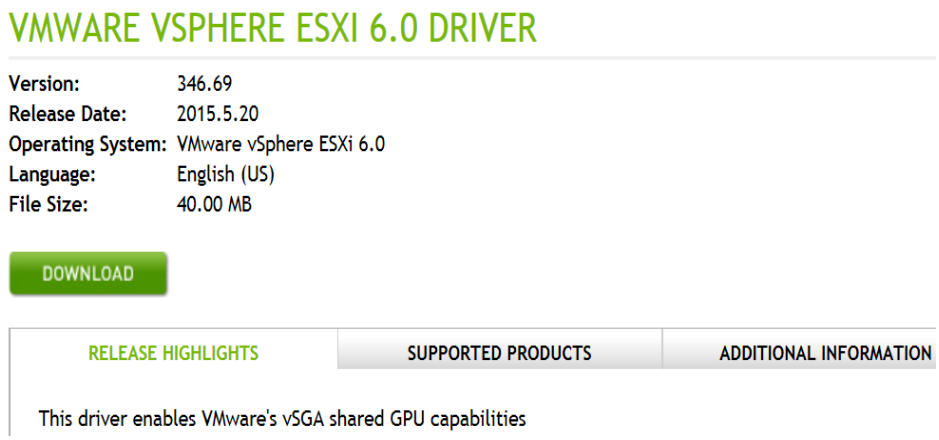


When you deploy vDGA, it uses the graphics driver from the GPU vendor rather than the virtual machine's vSGA 3D driver. To provide frame-buffer access, vDGA uses an interface between the remote protocol and graphics driver.

Configure vSGA GPU Deployment

1. Download the NVIDIA driver for vSphere ESXi 6.0 from http://www.nvidia.com/download/driverResults.aspx/85391/en-us (Figure 31).

Figure 31. Download the ESXi 6.0 Driver



2. Extract the downloaded file and install the vSphere Installation Bundle (VIB) on the ESXi host (shared storage is preferred if you are installing drivers on multiple servers) or using the VMware Update Manager.

Note: See the VMware documentation for information about installing the patch or host extension on ESXi.

Note: The ESXi host must be in maintenance mode for you to install the VIB module.

3. Download the NVIDIA VIB files and copy them to a local drive. You can install the files from the command line on the ESXi host using the command:

   esxcli software vib install –v /vib.file <path name>

```
[root@vGPU-01:~] esxcli software vib install -v /vmfs/volumes/datastore1\ \(4\)/NVIDIA-VMw
Host_Driver_346.42-1OEM.600.0.0.2159203.vib
Installation Result
   Message: Operation finished successfully.
   Reboot Required: false
   VIBs Installed: NVIDIA_bootbank_NVIDIA-VMware_ESXi_6.0_Host_Driver_346.42-1OEM.600.0.0.2
159203
   VIBs Removed:
   VIBs Skipped:
```

4. Verify that the the drivers were successfully installed on the ESXi host as shown here.

   esxcli software vib list | grep NVIDIA

```
[root@vGPU-01:~] esxcli software vib list | grep NVIDIA
NVIDIA-VMware_ESXi_6.0_Host_Driver  346.42-1OEM.600.0.0.2159203          NVIDIA  VMwareAc
cepted    2015-05-19
[root@vGPU-01:~]
```

Note: See the VMware knowledge base article for information about removing any existing NVIDIA drivers before installing new drivers for any vSGA or vDGA test: http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2033434.

5. Check the GRID cards installed from the command line on the ESXi host. The output shows the mode of the GPU card being configured and the amount of memory available on the card. Enter the following command:

   gpuvm

```
[root@vGPU-01:~] gpuvm
Xserver unix:0, PCI ID 0:11:0:0, vSGA mode, GPU maximum memory 4173824KB
        GPU memory left 4173824KB.
Xserver unix:1, PCI ID 0:12:0:0, vSGA mode, GPU maximum memory 4173824KB
        GPU memory left 4173824KB.
Xserver unix:2, PCI ID 0:134:0:0, vSGA mode, GPU maximum memory 4173824KB
        GPU memory left 4173824KB.
Xserver unix:3, PCI ID 0:135:0:0, vSGA mode, GPU maximum memory 4173824KB
        GPU memory left 4173824KB.
```

6. Query the status of the GPU card's CPU, the card's memory, and disk space left on the card by entering the following command:

   nvidia-smi

```
[root@vGPU-01:~] nvidia-smi
Tue May 19 20:41:07 2015
+------------------------------------------------------+
| NVIDIA-SMI 346.42     Driver Version: 346.42         |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  GRID K2             Off  | 0000:0B:00.0     Off |                  Off |
| N/A   47C    P8    17W / 117W |    107MiB /  4095MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   1  GRID K2             Off  | 0000:0C:00.0     Off |                  Off |
| N/A   41C    P8    17W / 117W |     89MiB /  4095MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   2  GRID K2             Off  | 0000:86:00.0     Off |                  Off |
| N/A   48C    P8    17W / 117W |     92MiB /  4095MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   3  GRID K2             Off  | 0000:87:00.0     Off |                  Off |
| N/A   41C    P8    17W / 117W |     92MiB /  4095MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID  Type  Process name                               Usage      |
|=============================================================================|
|    0     35520    G   Xorg                                           6MiB |
|    0     41145    G   vSGA-3                                         3MiB |
|    0     41146    G   vSGA-2                                        75MiB |
|    0     41222    G   vSGA-4                                         3MiB |
|    0     41304    G   vSGA-1                                         3MiB |
|    1     35674    G   Xorg                                           6MiB |
|    1     41145    G   vSGA-3                                        70MiB |
|    2     35948    G   Xorg                                           6MiB |
|    2     41222    G   vSGA-4                                        73MiB |
|    3     36100    G   Xorg                                           6MiB |
|    3     41304    G   vSGA-1                                        73MiB |
+-----------------------------------------------------------------------------+
[root@vGPU-01:~]
```

7. Start Xorg services on the host. You can either use the command line through SSH on the ESXi host by entering the command **/etc/init.d/Xorg status** or using vSphere Web Client.

   #SSH into ESXI Host

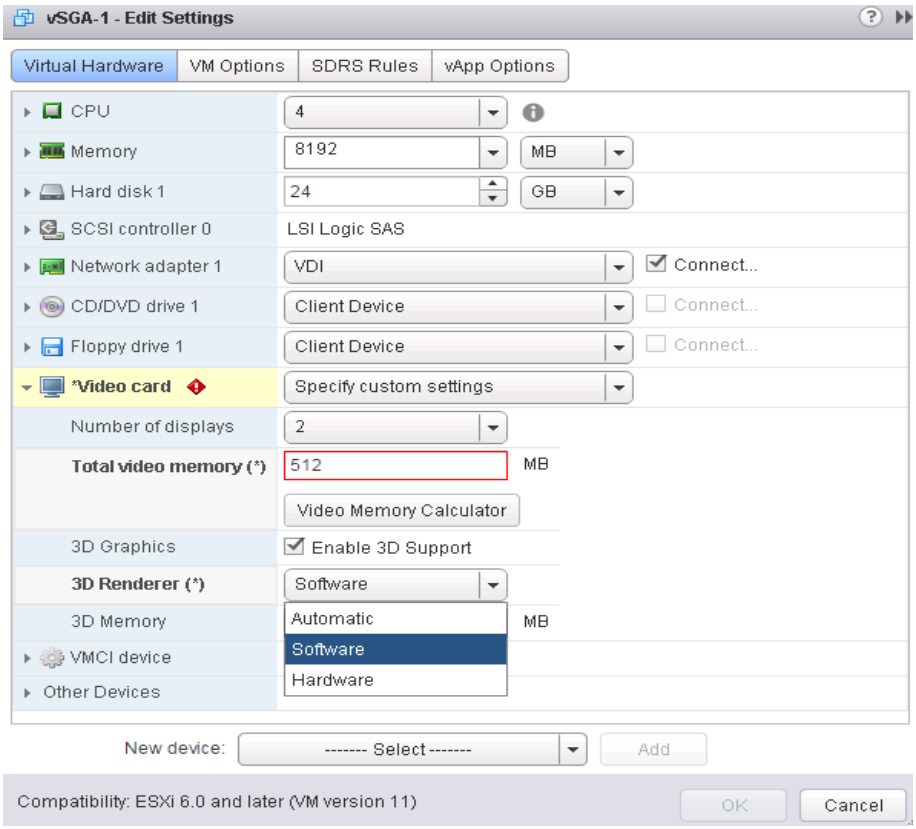
   /etc/init.d/Xorg start

   /etc/init.d/Xorg stop

```
[root@vSGA-01:~] /etc/init.d/xorg status
Xorg is not running
[root@vSGA-01:~] /etc/init.d/xorg start
Xorg0 started
Xorg1 started
Xorg2 started
Xorg3 started
[root@vSGA-01:~] /etc/init.d/xorg status
Xorg is running
```

For more information, see
http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2064775.

Figure 32. Editing Virtual Machine Properties for vSGA



Configure 3D Rendering Using VMware vSphere

When you configure vSGA and vDGA through the vSphere Web interface, you can choose from four 3D rendering options:

- Automatic uses hardware acceleration if the host in which the virtual machine is running contains a capable and available hardware GPU. If a hardware GPU is not available, the Automatic option uses software 3D rendering for any 3D tasks. This option allows the virtual machine to be started on, or migrated to (through vMotion), any host (vSphere 5.0 or later) and uses the best graphics solution available on that host.
- Software Only uses vSphere software 3D rendering, even if a hardware GPU is available on the host in which the virtual machine is running. This option does not provide the performance benefits that hardware 3D acceleration offers. However, this configuration allows the virtual machine to run on any host (vSphere 5.0 or later) and allows you to block virtual machines from using a hardware GPU in a host.
- Hardware Only uses hardware-accelerated GPUs. If a hardware GPU is not present in a host, either the virtual machine will not start or you will not be able to perform live vMotion migration of the virtual machine to that host. As long as the host to which the virtual machine is being moved has a capable and available hardware GPU, vMotion is possible with this option. You can use this option to help ensure that a virtual

machine will always use hardware 3D rendering when a GPU is available, but this option also limits the virtual machine to hosts with hardware GPUs.

- Disabled does not use 3D rendering at all (software or hardware) and overrides vSphere 3D settings to disable 3D rendering. Use this setting to help ensure that Horizon View desktop pools with nongraphical workloads do not use unnecessary resources: for instance, that they don't share a hardware GPU when running on the same cluster as Horizon View desktops with heavier graphics workloads.

For more information, see http://www.nvidia.com/download/driverResults.aspx/85390/en-us.

Configure vGPU Deployment

GRID vGPU allows multiple virtual desktops to share a single physical GPU, and multiple GPUs to reside on a single physical PCI card, all providing the 100 percent application compatibility of vDGA pass-through graphics, but with lower cost because multiple desktops share a single graphics card. With Horizon, you can centralize, pool, and more easily manage traditionally complex and expensive distributed workstations and desktops. Now all your user groups can take advantage of the benefits of virtualization.

NVIDIA GRID vGPU brings the full benefits of NVIDIA hardware-accelerated graphics to virtualized solutions. This technology provides exceptional graphics performance for virtual desktops equivalent to local PCs that share a GPU among multiple users.

GRID vGPU is the industry's most advanced technology for sharing true GPU hardware acceleration between multiple virtual desktops—without compromising the graphics experience. Application features and compatibility are exactly the same as they would be at the user's desk.

With GRID vGPU technology, the graphics commands of each virtual machine are passed directly to the GPU, without translation by the hypervisor. By allowing multiple virtual machines to access the power of a single GPU within the virtualization server, enterprises can increase the number of users with access to true GPU-based graphics acceleration on virtual machines.

The physical GPU within the server can be configured with a specific vGPU profile. Organizations have a great deal of flexibility in how best to configure their servers to meet the needs of various types of end users.

vGPU support allows businesses to use the power of NVIDIA's GRID technology to create a whole new class of virtual machines designed to provide end users with a rich, interactive graphics experience.

vGPU Profiles

Within any given enterprise, the needs of individual users vary widely. One of the main benefits of GRID vGPU is the flexibility to use various vGPU profiles designed to serve the needs of different classes of end users.

Although the needs of end users can be diverse, for simplicity users can be grouped into the following categories: knowledge workers, designers, and power users.

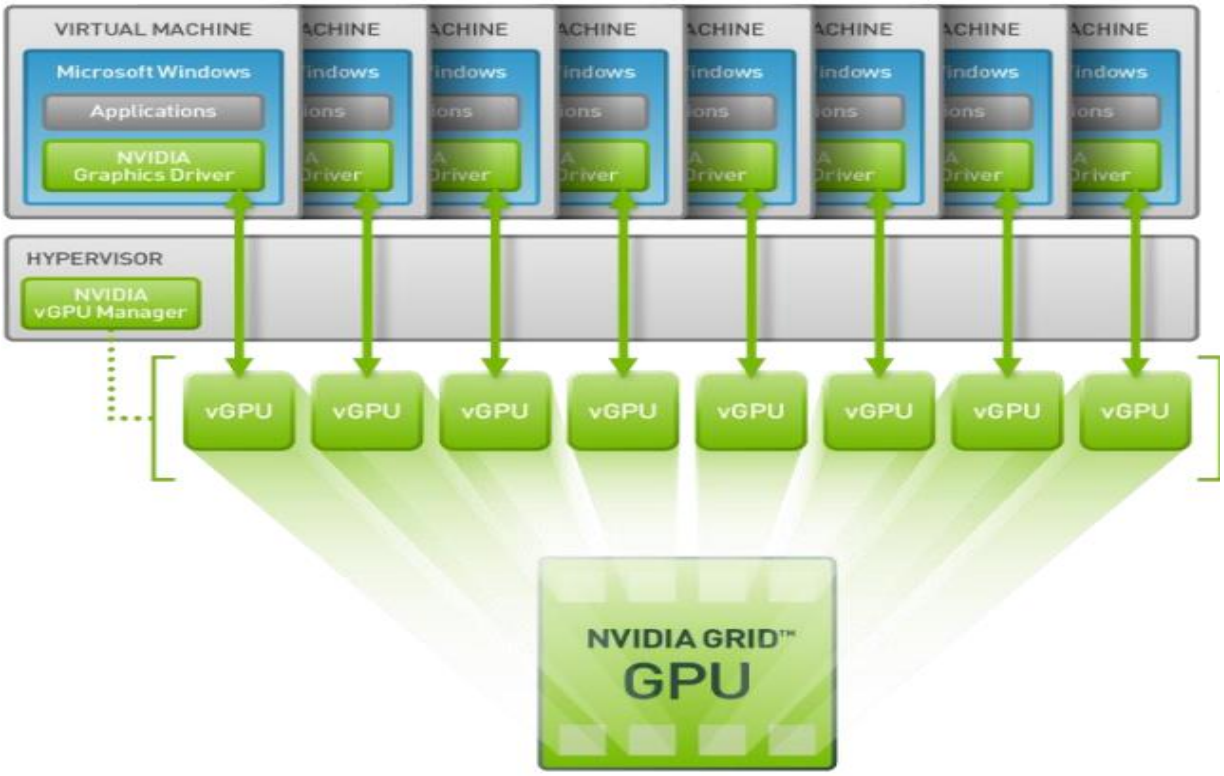
- For knowledge workers, the main areas of importance include office productivity applications, a rich web experience, and fluid video playback. Knowledge workers have the least-intensive graphics demands, but they expect the same smooth, fluid experience that exists natively on today's graphics accelerated devices such as desktop PCs, notebooks, tablets, and smartphones.
- Power users are users who need to run more demanding office applications, such as office productivity software, image editing software such as Adobe Photoshop, mainstream CAD software such as Autodesk

AutoCAD, and product lifecycle management (PLM) applications. These applications are more demanding and require additional graphics resources with full support for APIs such as OpenGL and Direct3D.

- Designers are users within an organization who run demanding professional applications such as high-end CAD software and professional digital content creation (DCC) tools. Examples include Autodesk Inventor, PTC Creo, Autodesk Revit, and Adobe Premiere. Historically, designers have used desktop workstations and have been a difficult group to incorporate into virtual deployments due to their need for high-end graphics and the certification requirements of professional CAD and DCC software.

vGPU profiles allow the GPU hardware to be time-sliced to deliver exceptional shared virtualized graphics performance (Figure 33).

Figure 33. vGPU GPU Architecture



Download vGPU Drivers

To get started with vGPU deployment, download the ESXi drivers from the NVIDIA website: http://www.nvidia.com/download/driverResults.aspx/85390/en-us.

The download will contain a package of drivers for the ESXi host and NVIDIA graphics software drivers for virtual machines (Figure 34).

Figure 34. Download vGPU Drivers

Install ESXi Host Drivers

1. After downloading the ESXi drivers, install them from the command line using SSH on the ESXi host and run the esxcli command shown here:

   esxcli software vib install –v /vib.file <path name>

```
[root@GPU-01:/vmfs/volumes/55365ea1-5bb3c32b-fa7f-0025b5041503] cd //
[root@GPU-01:~] esxcli software vib install -v /vmfs/volumes/datastore1\ \(4\)/NVIDIA-vgx-VMware_ESXi_6.0_Host_Driver_346
.42-1OEM.600.0.0.2159203.vib
Installation Result
   Message: Operation finished successfully.
   Reboot Required: false
   VIBs Installed: NVIDIA_bootbank_NVIDIA-vgx-VMware_ESXi_6.0_Host_Driver_346.42-1OEM.600.0.0.2159203
   VIBs Removed:
   VIBs Skipped:
[root@GPU-01:~]
```

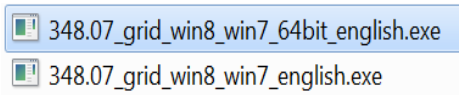
2. Verify that the drivers were successfully installed on the ESXi host by entering the following command:

   esxcli software vib list | grep NVIDIA

Install NVIDIA Virtual Machine Graphics Drivers

1. Choose the correct version of the virtual machine graphics drivers for the guest OS: either 32-bit or 64-bit (Figure 35).

Figure 35. Choosing the Correct Driver Version



2. Install the NVIDIA virtual machine graphics drivers (Figure 36 and 37).

Figure 36. Install NVIDIA Graphics Drivers on the Virtual Machine

Figure 37. Virtual Machine Drivers for vGPU Deployment

GRID K1 and K2 Profile Specifications

The GRID vGPU allows up to eight users to share each physical GPU, assigning the graphics resources of the available GPUs to virtual machines using a balanced approach. Each GRID K1 card has four GPUs, allowing 32 users to share a single card. Each GRID K2 card has two GPUs, allowing 16 users to share a single card. Table 8 summarizes the user profile specifications

For more information, see http://www.nvidia.com/object/virtual-gpus.html - sthash.aOZ68uVk.dpuf.

Table 8. User Profile Specifications for GRID K1 and K2 Cards

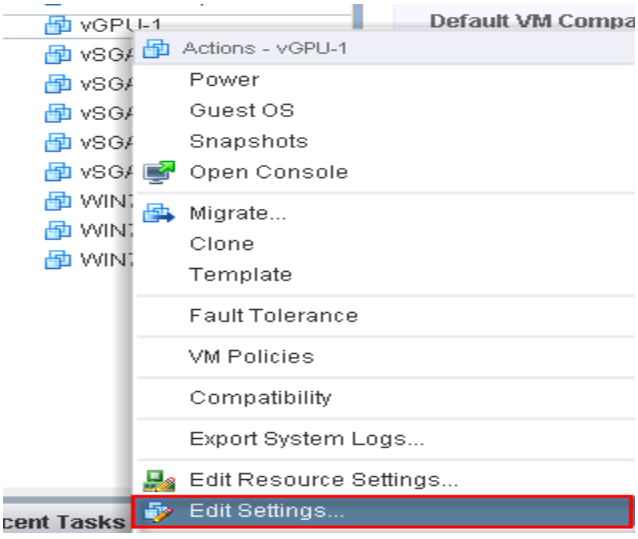
NVIDIA GRID CARD	Virtual GPU Profile	Application Certification	Graphics Memory in MB	Max Display per User	Max Resolution per Display	Max Users per Board	Use Case
GRID K2	K280Q	YES	4096	4	2560x1600	2	Designer
	K260Q	YES	2048	4	2560x1600	4	Designer/ Power User
	K240Q	YES	1024	2	2560x1600	8	Designer/ Power User
	K220Q	YES	512	2	2560x1600	16	Power User
GRID K1	K180Q	YES	4096	4	2560x1600	4	Power User
	K160Q	YES	2048	4	2560x1600	8	Power User
	K140Q	YES	1024	2	2560x1600	16	Knowledge Worker
	K120Q	YES	512	2	2560x1600	32	Knowledge Worker

Attach Virtual Machine Profiles for Users or Virtual Machines

Attach a profile for a user of virtual machine to match the user profile or workload scenario.

1. Log in to the vSphere Web Client and select the virtual machine properties. Choose Edit Settings (Figure 38).

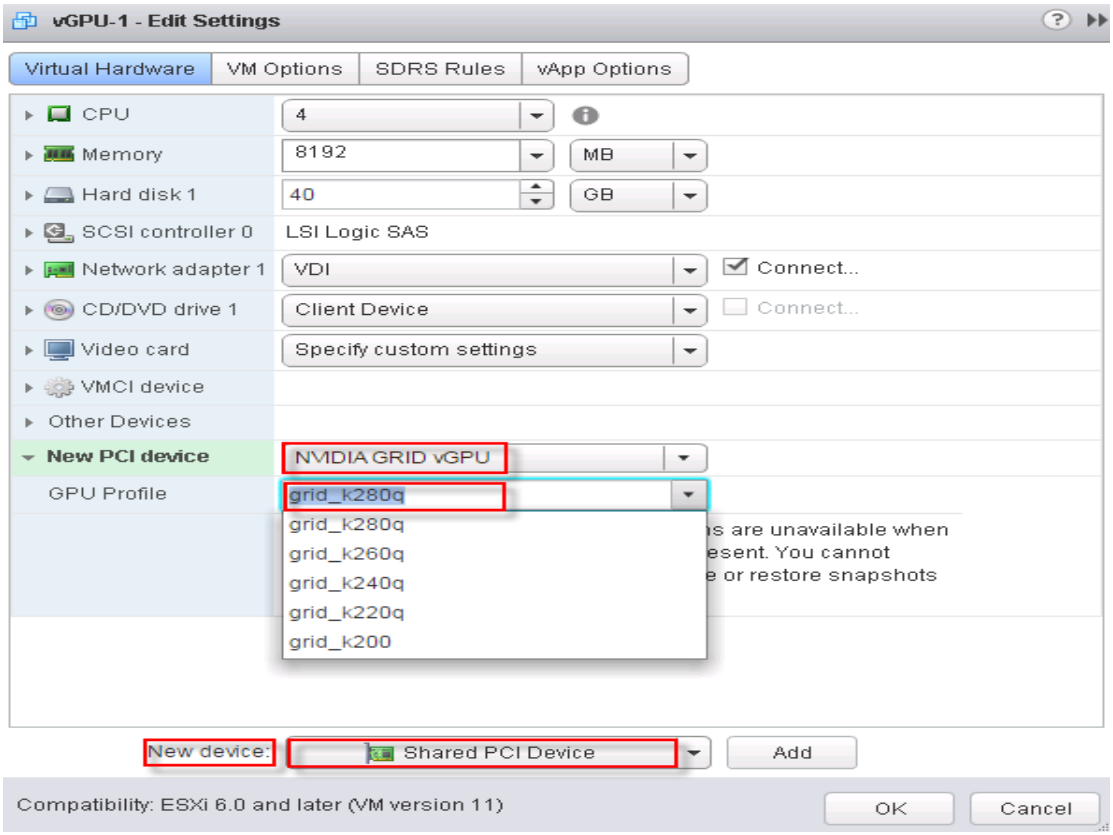
Figure 38. Editing Virtual Machine Properties to Add a GPU Profile to the Virtual Machine



2. Select the shared PCI device from the “New PCI device” menu (Figure 39). Then click Add.

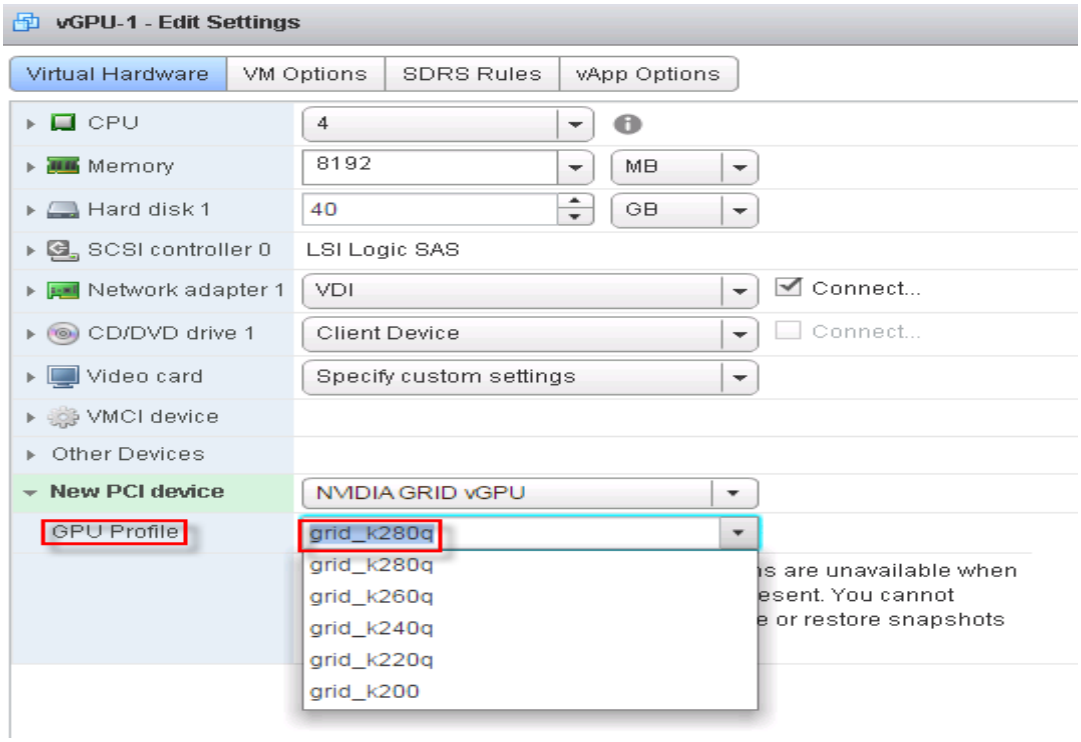
Note: Make sure that the virtual machine is shut down before you attach a shared PCI device.

Figure 39. Adding a Shared PCI Device from the New Device List



3. Attach a profile (Figure 40).

Figure 40. Attaching GPU Profile K280q to the Virtual Machine



Verify the Attached GPU Profile from the Command Line

You can also use the command line to check the NVIDIA grid cards installed on the ESXI host. The output shows the mode of the GPU card that is being configured and the amount of memory available on the card. Enter this command:

gpuvm

```
[root@vGPU-01:~] gpuvm
Xserver unix:0, PCI ID 0:11:0:0, vGPU: 0x11b0:0x113d, GPU maximum memory 4182864KB
        pid 38329, VM "vGPU-1", reserved 3866624KB of GPU memory.
        GPU memory left 316240KB.
Xserver unix:1, PCI ID 0:12:0:0, vGPU: Not set, GPU maximum memory 4182864KB
        GPU memory left 4182864KB.
Xserver unix:2, PCI ID 0:134:0:0, vGPU: Not set, GPU maximum memory 4182864KB
        GPU memory left 4182864KB.
Xserver unix:3, PCI ID 0:135:0:0, vGPU: Not set, GPU maximum memory 4182864KB
        GPU memory left 4182864KB.
[root@vGPU-01:~]
```

Use the **lspci** command to display all the NVIDIA GRID card VGA controllers present on the ESXi host
lspci | grep –i display

```
[root@vGPU-01:~] lspci | grep -i display
0000:0b:00.0 Display controller: NVIDIA Corporation NVIDIAGRID K2 [vmgfx0]
0000:0c:00.0 Display controller: NVIDIA Corporation NVIDIAGRID K2 [vmgfx1]
0000:0e:00.0 Display controller: Matrox Electronics Systems Ltd. MGA G200e
0000:86:00.0 Display controller: NVIDIA Corporation NVIDIAGRID K2 [vmgfx2]
0000:87:00.0 Display controller: NVIDIA Corporation NVIDIAGRID K2 [vmgfx3]
[root@vGPU-01:~]
```

Configure 3D Rendering Using VMware Horizon View

1. Configure vSGA and vDGA in the Horizon View 6.1 Desktop Pool settings using one of the five 3D rendering options:
   - Manage Using vSphere Client will not make any changes to the 3D settings of the individual virtual machines in that pool. This option allows individual virtual machines to have different settings set through vSphere. You will most likely use this setting during testing or for manual desktop pools.
   - Automatic uses hardware acceleration if the host in which the virtual machine is running contains a capable and available hardware GPU. If a hardware GPU is not available, this option uses software 3D rendering for any 3D tasks. This option allows the virtual machine to be started on, or migrated to (through vMotion), any host (vSphere 5.0 or later) and uses the best solution available on that host.
   - Software Only uses vSphere software 3D rendering, even if the host in which the virtual machine is running has an available hardware GPU. This option does not provide the performance benefits that hardware 3D acceleration offers. However, it allows the virtual machine to run on any host (vSphere 5.0 or later) and allows you to block virtual machines from using hardware GPU on a host.
   - Hardware Only uses hardware-accelerated GPUs. If a hardware GPU is not present in a host, either the virtual machine will not start or you will not be able to perform live vMotion migration of the virtual machine to that host. As long as the host to which the virtual machine is being moved has a capable and available hardware GPU, vMotion is possible with this option. You can use this option to help ensure that a virtual machine will always use hardware 3D rendering when a GPU is available; but this option limits the virtual machine to hosts with hardware GPUs.
   - Disabled does not use 3D rendering at all (software or hardware) and overrides vSphere 3D settings to disable 3D rendering. Use this setting to help ensure that Horizon View desktop pools with nongraphical workloads do not use unnecessary resources: for instance, that they don't share a hardware GPU when running on the same cluster as Horizon View desktops with heavier graphics workloads.

2. For the discussion here, PCoIP is used as the default display protocol, and users are not allowed to choose the display protocol. Select Hardware from the drop-down menu for the 3D renderer and click Configure to select the VRAM (video memory) size. Table 9 lists the VRAM ranges available.

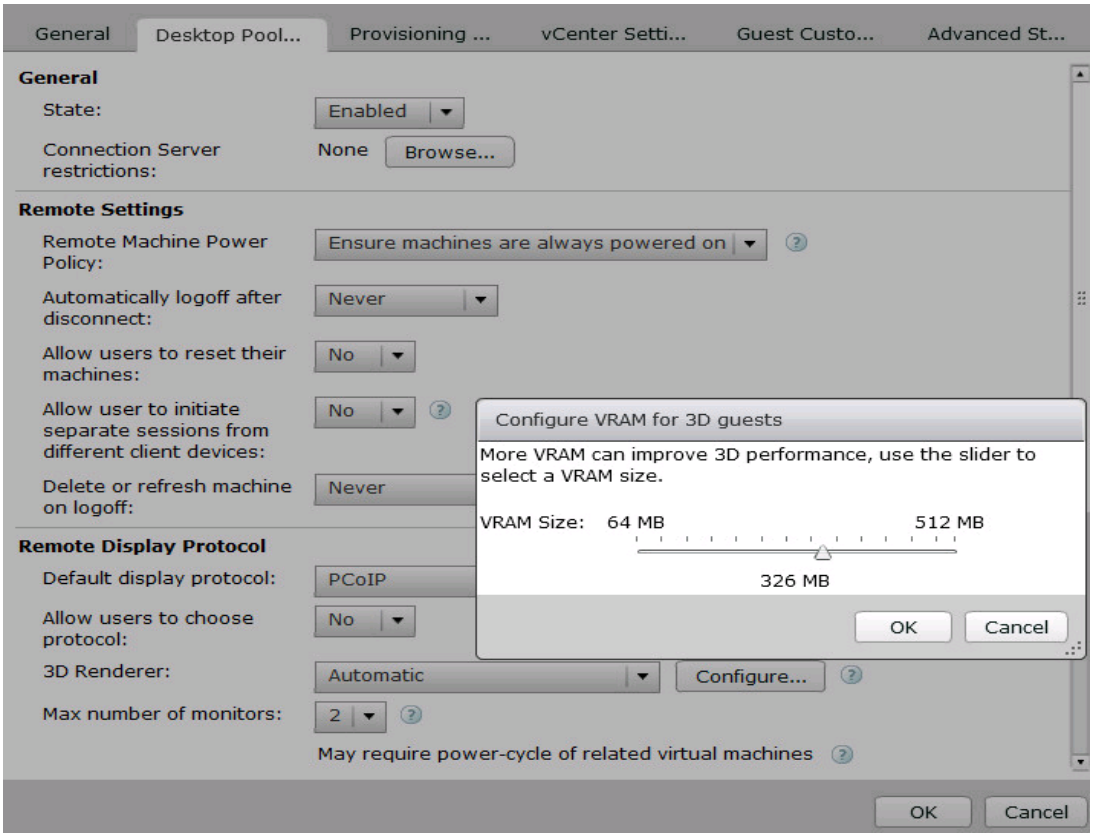
Table 9. Minimum and Maximum VRAM Sizes for vSGA and vDGA

	Soft 3D (Software 3D)	vSGA (Hardware 3D)
Minimum	1.18 MB	64 MB
Default	64 MB	96 MB
Maximum	512 MB	512 MB

Note the following when configuring VRAM settings:

- Whenever the 3D renderer setting changes, the amount of video memory reserved is the 96-MB default. Make sure that you change the video memory to the appropriate amount after you change this setting (Figure 41).

Figure 41. Configuring VRAM for 3D Rendering from the Horizon View Administrator Console



- VRAM settings that you configure in the Horizon View Administrator take precedence over VRAM settings that are configured for the virtual machines in the vSphere Client or vSphere Web Client. Select the Manage Using vSphere Client option to prevent this behavior.
- If you are using the Manage Using vSphere Client option, VMware recommends that you use the web client to configure the virtual machines, rather than the traditional vSphere Client because the traditional vSphere Client does not display the various rendering options; it displays only Enable or Disable 3D support.

3. After you make VRAM changes to the Horizon View pool, a short delay may occur (sometimes a few minutes) before the message “Reconfiguring virtual machine” appears in the vCenter console. Be sure to wait for this process to complete before power-cycling the virtual machines.
4. If you enable the 3D renderer settings, configure the maximum number of monitors. You can specify one or two monitors. You cannot select more than two monitors. The maximum resolution of any one monitor setting is 1920 x 1200 pixels. You cannot configure a higher value.

Note: You must power existing virtual machines off and then on for the 3D renderer setting to take effect. Restarting or rebooting a virtual machine does not cause the setting to take effect.

Note: Note: This example scenario used Hardware Only 3D rendering with 512 MB of VRAM per virtual machine for vSGA. For vDGA, the Manage Using vSphere Client option was used, with 512 MB of VRAM.

Performance Tuning Tips

This section offers tips to improve the performance of virtual machines for NVIDIA GPU deployment.

Configuring Adequate Virtual Machine Resources

Desktops that use high-end 3D capabilities must be provisioned with more vCPUs and memory than traditional VDI desktops. Make sure that your desktop virtual machines meet the memory and CPU requirements for the applications you use. The minimum requirements that VMware recommends for 3D workloads are two vCPUs and 4 GB of RAM.

Optimizing PCoIP

Occasionally, PCoIP custom configurations can contribute to poor performance. By default, PCoIP is set to allow a maximum of 30 frames per second (fps). Some applications require significantly more than that. If you notice that the frame rate of an application is lower than expected, reconfigure the PCoIP group policy object (GPO) settings to allow a maximum of 120 fps.

You can also disable the PCoIP Build-to-Lossless feature. This setting reduces the overall amount of PCoIP traffic, which reduces the load placed on both the virtual machine and the endpoint.

Enabling Relative Mouse

If you are using an application or game in which the cursor is moving uncontrollably, enabling the relative mouse feature may improve mouse control.

The relative mouse is a new feature in the Horizon View Client for Windows that changes the way client mouse movement is tracked and sent to the server through PCoIP. Traditionally, PCoIP uses absolute coordinates. Absolute mouse events allow the client to render the cursor locally, which is a significant optimization for high-latency environments. However, not all applications work well when using the absolute mouse. Two notable classes of applications, CAD applications and 3D games, rely on relative mouse events to function correctly.

With the introduction of vSGA and vDGA, VMware expects the requirements for relative mouse to increase rapidly as CAD and 3D games become more heavily used in Horizon View environments.

The end user can enable the relative mouse feature manually. To manually enable this feature, right-click the Horizon View Client shade at the top of the screen and select Relative Mouse. A check mark appears next to Relative Mouse.

Note: The Horizon View Client for Windows is required to enable the relative mouse feature. As of this writing, this feature is not available through any other Horizon View Clients or zero clients. Relative Mouse must be selected on each and every connection. Currently, no option is available to enable this feature by default.

Improving Performance for Virtual Machines Using VMXNET3

For desktop virtual machines using VMXNET3 Ethernet adapters, you can significantly improve peak video-playback performance of your Horizon View desktop by following these steps, which are recommended by Microsoft for virtual machines:

1. Start the Registry Editor (Regedt32.exe).
2. Locate the following key in the registry: KLM\System\CurrentControlSet\Services\Afd\Parameters.
3. On the Edit menu, choose Add Value and add the following registry entries:

   Value Name: FastSendDatagramThreshold
   Data Type: REG_DWORD
   Value: 1500

4. Quit the Registry Editor.

Note: You must reboot the desktop virtual machine after changing this registry setting. If this setting does not exist, create it as a DWORD value. For more information about this setting, see the Microsoft Support website.

Workaround for CAD Performance Problem
VMware has experienced a performance problem when users deploy Dassault Systèmes Computer Aided Three-dimensional Interactive Application (CATIA). Occasionally, when working with CAD models (when turning and spinning), you may find that objects move irregularly and with a delay. However, the objects themselves are displayed clearly, without blurring.

The workaround in this case is to disable the MaxApp FrameRate registry entry. The registry key can be found at HKLM\Software\VMware, Inc.\VMware SVGA DevTap\MaxAppFrameRate.

Change this registry setting to **dword: 00000000**.

Note: If this registry key does not exist, this setting defaults to 30.

Important: This change can negatively affect other applications. Use this workaround with caution and only if you are experiencing the symptoms mentioned here.

Conclusion

The combination of Cisco UCS Manager, Cisco UCS C240 M4 Rack Servers, NVIDIA GRID K1 and K2 or Tesla cards using VMware vSphere ESXi 6.0, and VMware Horizon 6.1 provides a high-performance platform for virtualizing graphics-intensive applications.

By following the guidance in this document, our customers and partners can be assured that they are ready to host the growing list of graphics applications that are supported by our partners.

For More Information

- Cisco UCS C-Series Rack Servers
  - http://www.cisco.com/en/US/products/ps10265/
  - http://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-c240-m4-rack-server/index.html
  - http://www.cisco.com/c/en/us/products/interfaces-modules/ucs-virtual-interface-card-1227/index.html
  - http://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-c-series-rack-servers/index.html

- NVIDIA
  - http://www.nvidia.com/content/cloud-computing/pdf/nvidia-grid-datasheet-k1-k2.pdf
  - http://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-b-series-blade-servers/tesla_kseries_overview_lr.pdf
  - http://www.nvidia.com/content/grid/pdf/GRID_K1_BD-06633-001_v02.pdf
  - http://www.nvidia.com/content/grid/pdf/GRID_K2_BD-06580-001_v02.pdf
  - http://www.nvidia.com/content/tesla/pdf/tesla-kseries-overview-lr.pdf

- VMware Horizon View 6.1
  - https://www.vmware.com/support/horizon-view/doc/horizon-61-view-release-notes.html
  - https://www.vmware.com/support/horizon-view/doc/horizon-view-602-release-notes.htmlhttps://pubs.vmware.com/horizon-view-60/index.jsp
  - http://blogs.vmware.com/performance/2014/12/vmware-horizon-6-hardware-accelerated-3d-graphics.html
  - http://www.vmware.com/files/pdf/techpaper/vmware-horizon-view-graphics-acceleration-deployment.pdf
  - http://www.vmware.com/files/pdf/view/vmware-horizon-view-best-practices-performance-study.pdf
  - http://www.vmware.com/files/pdf/products/horizon/vmware-nvidia-grid-vgpu-solution-brief.pdf
  - http://www.vmware.com/files/pdf/products/horizon/vmware-nvidia-grid-vgpu-FAQ.pdf

- VMware Horizon View 6 with PCoIP network optimization
  - http://blogs.vmware.com/euc/2014/06/vmware-horizon-view-6-pcoip-optimization-bandwidth-changes.html

- VMware Horizon View virtual desktop: Microsoft Windows 7 optimization
  - http://www.vmware.com/files/pdf/VMware-View-OptimizationGuideWindows7-EN.pdf

- VMware vSphere ESXi and vCenter Server 6
  - http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2107948
  - http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2109712
  - http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2033434

About the Author

Ramesh Guduru, Engineer/Developer, Cisco Systems Inc.

Ramesh Guduru, Virtualization System Engineer in Computer Systems Products Group, UCS Product Management and DC Solutions Engineering, Cisco Systems Inc.

Acknowledgements

Steve Harpester, NVIDIA, Inc.

Mike Brennan, Cisco Systems, Inc.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

Printed in USA C11-735450-00 07/15